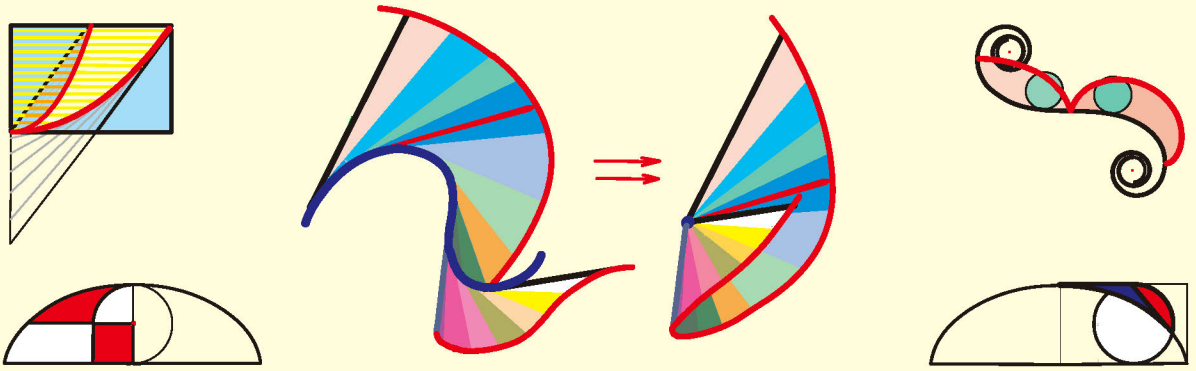
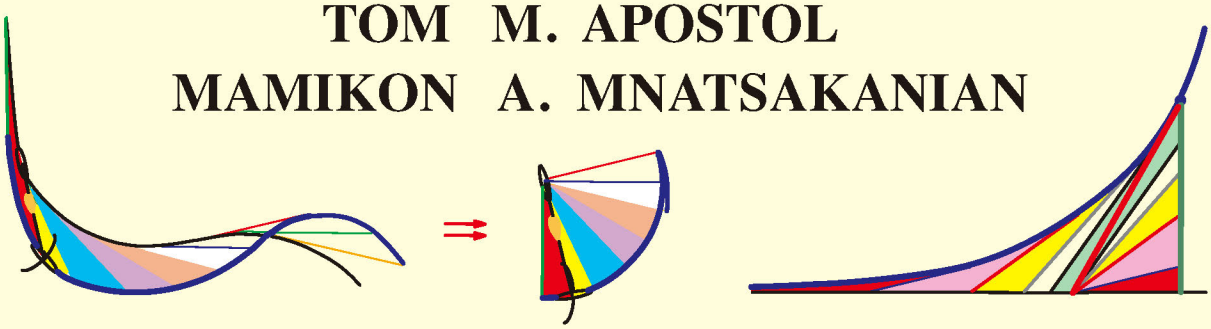
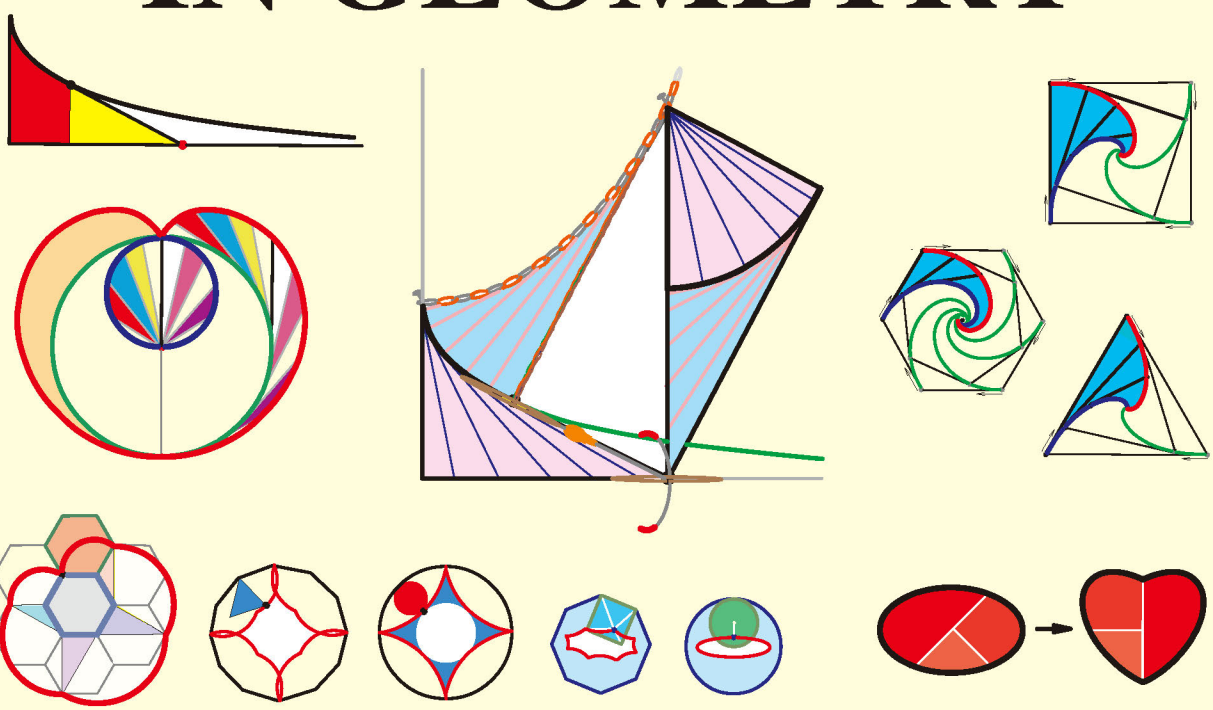


TOM M. APOSTOL
MAMIKON A. MNATSAKANIAN



NEW HORIZONS IN GEOMETRY



MAA MATHEMATICAL ASSOCIATION OF AMERICA
DOLCIANI MATHEMATICAL EXPOSITIONS # 47

NEW HORIZONS IN GEOMETRY

Library of Congress Catalog Card Number 2012949754

Print Edition ISBN 978-0-88385-354-2

Electronic Edition ISBN 978-1-61444-210-3

Printed in South Korea
by Charles Allen Imaging Experts, Pasadena, CA

Current Printing (last digit):
10 9 8 7 6 5 4 3 2 1

The Dolciani Mathematical Expositions
NUMBER FORTY-SEVEN

NEW HORIZONS IN GEOMETRY

Tom M. Apostol
California Institute of Technology

and

Mamikon A. Mnatsakanian
California Institute of Technology

Published and Distributed by
The Mathematical Association of America

The DOLCIANI MATHEMATICAL EXPOSITIONS series of the Mathematical Association of America was established through a generous gift to the Association from Mary P. Dolciani, Professor of Mathematics at Hunter College of the City University of New York. In making the gift, Professor Dolciani, herself an exceptionally talented and successful expositor of mathematics, had the purpose of furthering the ideal of excellence in mathematical exposition.

The Association, for its part, was delighted to accept the gracious gesture initiating the revolving fund for this series from one who has served the Association with distinction, both as a member of the Committee on Publications and as a member of the Board of Governors. It was with genuine pleasure that the Board chose to name the series in her honor.

The books in the series are selected for their lucid expository style and stimulating mathematical content. Typically, they contain an ample supply of exercises, many with accompanying solutions. They are intended to be sufficiently elementary for the undergraduate and even the mathematically inclined high-school student to understand and enjoy, but also to be interesting and sometimes challenging to the more advanced mathematician.

1. *Mathematical Gems,* Ross Honsberger
2. *Mathematical Gems II,* Ross Honsberger
3. *Mathematical Morsels,* Ross Honsberger
4. *Mathematical Plums,* Ross Honsberger (ed.)
5. *Great Moments in Mathematics* (*Before 1650*), Howard Eves
6. *Maxima and Minima without Calculus,* Ivan Niven
7. *Great Moments in Mathematics* (*After 1650*), Howard Eves
8. *Map Coloring, Polyhedra, and the Four-Color Problem,* David Barnette
9. *Mathematical Gems III,* Ross Honsberger
10. *More Mathematical Morsels,* Ross Honsberger
11. *Old and New Unsolved Problems in Plane Geometry and Number Theory,* Victor Klee and Stan Wagon
12. *Problems for Mathematicians, Young and Old,* Paul R. Halmos
13. *Excursions in Calculus: An Interplay of the Continuous and the Discrete,* Robert M. Young
14. *The Wohascum County Problem Book,* George T. Gilbert, Mark Krusemeyer, and Loren C. Larson
15. *Lion Hunting and Other Mathematical Pursuits: A Collection of Mathematics, Verse, and Stories by Ralph P. Boas, Jr.,* edited by Gerald L. Alexanderson and Dale H. Mugler
16. *Linear Algebra Problem Book,* Paul R. Halmos
17. *From Erdős to Kiev: Problems of Olympiad Caliber,* Ross Honsberger
18. *Which Way Did the Bicycle Go? . . . and Other Intriguing Mathematical Mysteries,* Joseph D. E. Konhauser, Dan Velleman, and Stan Wagon
19. *In Pólya's Footsteps: Miscellaneous Problems and Essays,* Ross Honsberger
20. *Diophantus and Diophantine Equations,* I. G. Bashmakova (Updated by Joseph Silverman and translated by Abe Shenitzer)
21. *Logic as Algebra,* Paul Halmos and Steven Givant
22. *Euler: The Master of Us All,* William Dunham
23. *The Beginnings and Evolution of Algebra,* I. G. Bashmakova and G. S. Smirnova (Translated by Abe Shenitzer)
24. *Mathematical Chestnuts from Around the World,* Ross Honsberger
25. *Counting on Frameworks: Mathematics to Aid the Design of Rigid Structures,* Jack E. Graver
26. *Mathematical Diamonds,* Ross Honsberger
27. *Proofs that Really Count: The Art of Combinatorial Proof,* Arthur T. Benjamin and Jennifer J. Quinn
28. *Mathematical Delights,* Ross Honsberger
29. *Conics,* Keith Kendig
30. *Hesiod's Anvil: falling and spinning through heaven and earth,* Andrew J. Simoson
31. *A Garden of Integrals,* Frank E. Burk
32. *A Guide to Complex Variables* (MAA Guides #1), Steven G. Krantz
33. *Sink or Float? Thought Problems in Math and Physics*, Keith Kendig
34. *Biscuits of Number Theory*, Arthur T. Benjamin and Ezra Brown
35. *Uncommon Mathematical Excursions: Polynomia and Related Realms*, Dan Kalman
36. *When Less is More: Visualizing Basic Inequalities*, Claudi Alsina and Roger B. Nelsen

37. *A Guide to Advanced Real Analysis* (MAA Guides #2), Gerald B. Folland
38. *A Guide to Real Variables* (MAA Guides #3), Steven G. Krantz
39. *Voltaire's Riddle: Micromégas and the measure of all things*, Andrew J. Simoson
40. *A Guide to Topology*, (MAA Guides #4), Steven G. Krantz
41. *A Guide to Elementary Number Theory*, (MAA Guides #5), Underwood Dudley
42. *Charming Proofs: A Journey into Elegant Mathematics*, Claudi Alsina and Roger B. Nelsen
43. *Mathematics and Sports*, edited by Joseph A. Gallian
44. *A Guide to Advanced Linear Algebra*, (MAA Guides #6), Steven H. Weintraub
45. *Icons of Mathematics: An Exploration of Twenty Key Images*, Claudi Alsina and Roger B. Nelsen
46. *A Guide to Plane Algebraic Curves*, (MAA Guides #7), Keith Kendig
47. *New Horizons in Geometry,* Tom M. Apostol and Mamikon A. Mnatsakanian
48. *A Guide to Groups, Rings, and Fields,* (MAA Guides #8), Fernando Gouvêa

MAA Service Center
P.O. Box 91112
Washington, DC 20090-1112
1-800-331-1MAA FAX: 1-301-206-9789

CONTENTS*

*Detailed contents for each chapter appear at the beginning of the chapter

PREFACE

This book is a compendium of joint work produced by the authors during the period 1998–2012, most of it published in the *American Mathematical Monthly, Math Horizons, Mathematics Magazine,* and *The Mathematical Gazette.* The published papers have been edited, augmented, and rearranged into 15 chapters. Each chapter is preceded by a sample of problems that can be solved by the methods developed in that chapter. Each opening page contains a brief abstract of the chapter's contents.

Chapter 1, entitled "Mamikon's Sweeping-Tangent Theorem," was the starting point of this collaboration. It describes an innovative and visual approach for solving many standard calculus problems by a geometric method that makes little or no use of formulas. The method was conceived in 1959 by my co-author (who prefers to be called Mamikon), when he was an undergraduate student at Yerevan University in Armenia. When young Mamikon showed his method to Soviet mathematicians they dismissed it out of hand and said "It can't be right. You can't solve calculus problems that easily."

Mamikon went on to get a Ph.D. in physics, was appointed a professor of astrophysics at the University of Yerevan, and became an international expert in radiative transfer theory, all the while continuing to develop his powerful geometric methods. Mamikon eventually published a paper outlining them in 1981, but it seems to have escaped notice, probably because it appeared in Russian in an Armenian journal with limited circulation. (Reference [59] in the Bibliography.)

Mamikon came to California in 1990 to learn more about earthquake preparedness for Armenia. When the Soviet government collapsed he was stranded in the United States without a visa. With the help of a few mathematicians he had met in Sacramento and at UC Davis and who recognized his remarkable talents, Mamikon was granted status as an "alien of extraordinary ability." While working at UC Davis and for the California Department of Education, he further developed his methods into a universal teaching tool using hands-on and computer activities, as well as diagrams. He has taught these methods at UC Davis and in Northern California classrooms, ranging from Montessori elementary schools to inner-city public high schools, and he has demonstrated them at teacher conferences. Students and teachers alike have responded enthusiastically, because the methods are vivid and dynamic and don't require the algebraic formalism of trigonometry or calculus.

A few years later, Mamikon visited Caltech and convinced me that his methods have the potential to make a significant impact on mathematics education, especially if they are combined with visualization tools of modern technology. Since then, we

have jointly published thirty papers, not only on his sweeping-tangent method, but also on a variety of topics in mathematics that are amenable to Mamikon's remarkable geometric insight.

I have often described Mamikon as "an artesian well of ideas." It has been a pleasure to work with him in an effort to share many of these ideas with those who enjoy the beauty of mathematics.

Tom M. Apostol
Professor of Mathematics, Emeritus, California Institute of Technology

INTRODUCTION

Since 1997 Tom Apostol and Mamikon Mnatsakanian have co-authored 30 papers, most of which are on geometry. Their work is strikingly innovative, combining the classical with the modern. They often surprise the reader with fresh, frequently astounding conclusions that challenge the imagination. As an added attraction to the reader, only a modest background is needed to understand their work. Working together, they have won three Lester R. Ford Awards since 2005 for five papers published in the *American Mathematical Monthly* in 2004, 2007, and 2009 — an enviable achievement.

The citation for their 2004 Ford Award notes that they do classical geometry with a modern twist, and modern geometry with a classical twist, obtaining new and surprising results in areas that have been mined for centuries. Their writing is hailed as a model of the succinct and the elegant, and a rich mix of the new and the classical.

This book gathers together several of their papers that constitute a royal road through several parts of classical geometry with spectacular side trips down lanes that previously were not known. In this volume they have expanded on the original papers and added exercises. Their writing also brings new richness to calculus. Newton and Leibniz would likely have been grateful for their wonderfully intuitive insights. Mamikon's approach to many integration problems has great power and should be more widely known.

Apostol and Mnatsakanian breathe new life into classical geometry, and they frequently praise the old masters, especially Archimedes, who became a hero for Mamikon when he was a boy. It is hardly surprising that many have remarked that their work is in the spirit of Archimedes.

The story of the Apostol - Mamikon partnership is too long for this introduction, but it's safe to say that it is a remarkably productive collaboration. Over the past 13 years they have co-authored 30 papers, 14 in the *American Mathematical Monthly*! Apostol, a member of the Caltech faculty since 1950, is a distinguished and vigorous eighty-nine-year old number theorist and author of more than 100 papers and 61 books, including his famous two-volume *Calculus*, which has been translated into several languages, was published more than a half century ago, and is still in print. He also is the creator and director of *Project Mathematics!*, a prize-winning series of mathematics videos.

Mnatsakanian was a theoretical astrophysicist and professor at Yerevan University in Armenia for many years. In 1990 while working in California on an

earthquake-preparedness program for Armenia, he was stranded in the United States when the government of Armenia collapsed. Through an amazing sequence of events, he and Apostol became collaborators. He is the author of more than 80 papers. Mamikon's mathematical ability is complemented by his artistic talents; his freehand sketches often rival computer generated drawings. As he describes a new idea, he invariably draws a helpful picture that provides illumination. I have had the pleasure of visiting with them at Caltech for the past thirteen years, and on my visits I have regularly been treated to exciting ideas that they are developing.

Mnatsakanian and Apostol have devised new geometrical methods for solving a host of mathematical problems, and the first chapter of this book provides an introduction to Mamikon's theorem, a result of great power. His theorem has considerable intuitive appeal, and is easily understood, even by elementary school children. Many difficult problems from calculus and differential equations are dispatched with ease by use of his theorem. Take for example a unit tangent continuously moving around the perimeter of an ellipse with major axis 16 and minor axis 9. Find the area of the oval swept out by the moving tangent. His theorem makes use of a sweeping tangent, and underscores the utility of many clever dynamic arguments that they bring to first solve and then extend many classical problems of geometry. Read how they use Mamikon's theorem to solve what is generally regarded as a more difficult problem - finding the area under a tractrix and above the x-axis.

The extensions of established results found by Apostol and Mnatsakanian are often stunning and unexpected. For example, it is well known that the area under a cycloidal arch is three times the area of the generating circle. In Chapter 2 you will learn the not-so-well-known fact that the latter area relationship holds throughout the generation of the cycloid. That is, the area of the cycloidal sector at each instant of its generation is three times the area of the circular segment determined by the portion of the perimeter through which the circle has rolled.

In Chapter 13, they introduce new and powerful balancing principles, including double equilibrium. The reader may recall Archimedes' use of mechanical balancing methods to prove that the volume of a sphere is two-thirds that of the cylinder circumscribing it. Their new balancing principles yield not only the volume results of Archimedes but also a number of surprising relations involving both volumes and surface areas of circumsolids of revolution as well as higher-dimensional spheres, cylindroids, spherical wedges, and cylindrical wedges. Apostol and Mamikon adhere to Archimedes' style of reducing properties of complicated objects to those of simpler objects.

In the pages ahead you will encounter many new ideas: tangent sweep, tangent cluster, cyclogons, circumgons, circumsolids, Archimedean globes, and more. You're in for a great ride in the spirit of Archimedes through a beautiful geometrical landscape that will give you considerable pleasure and a heightened appreciation for a wonderful subject.

Don Albers
Director of Publications, Emeritus

FOREWORD

Mathematics is not alien and remote but just a very human exploration of the patterns of the world, one which thrives on play and surprise and beauty.
– Indra's Pearls: The Vision of Felix Klein

This passage perfectly captures the spirit of the book *New Horizons in Geometry*, by Tom Apostol and Mamikon Mnatsakanian. In a remarkable display of mathematical versatility and imagination, the authors present us with a wealth of geometrical gems. These beautiful and often surprising results deal with a multitude of geometric forms, their interrelationships, and in many cases, their connection with patterns underlying the laws of nature. Lengths, areas and volumes, of curves, surfaces, and solids, are explored from a visually captivating perspective. The preponderance of results discussed by the authors are new, and when not new, are presented with unusual insights and unexpected generalizations.

The exposition is uniformly lucid and delightful, with a heavy emphasis on dynamic visual thinking. Some derivations that might ordinarily be carried out using methods of calculus are accomplished with ingenious visual arguments. For instance, an amazing variety of results are derived visually for cycloids, epicycloids and hypocycloids, general roulettes, pursuit curves, traces and envelopes of trammels, conic sections, and so forth. Classical constructions and characterizations of the conics are generalized, with focal points replaced by focal disks, and Dandelin spheres inside cones replaced by tangent spheres inside twisted cylinders.

Constructions and mechanical interpretations in the spirit of Archimedes involving centroids and moments are carried to new heights and to higher dimensional spaces. Several results not amenable to formulation as calculus problems are obtained with elegant geometric methods. There are many engaging and adjunct geometric treatments of results traditionally approached via calculus. For example, Mamikon's Sweeping-Tangent Theorem and the tractrix as involute of the catenary cooperate to give a lovely "proof without words" that the length of an arc of the catenary is proportional to the area under the arc. Placing these striking geometric representations alongside their analytic counterparts in calculus reveals that mathematics has an aesthetic dimension that can serve as a propelling force behind the human exploration referred to above in the quote from Indra's Pearls. Seeing standard calculus topics enriched with examples from these chapters can enhance the student's understanding and bring deeper meaning to these topics. We are reminded that the artistic impulse and intuition carry along, support, and mo-

tivate scientific expertise and technical skill in exploring the world of patterns and applying our discoveries to the patterns of the world. The student will acquire a vision of mathematics beyond that of a static formal system and will sense that it is a structure of great beauty, intriguing, dynamic and multifaceted, the elements of which can be found embodied in nature. This book should be kept at hand by instructors in geometry and calculus courses, to provide supplementary material that will enchant and inspire both students and teachers.

Don Chakerian
Professor of Mathematics, Emeritus, University of California, Davis

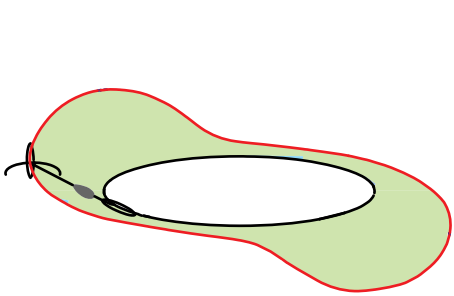
Chapter 1

MAMIKON'S SWEEPING-TANGENT THEOREM

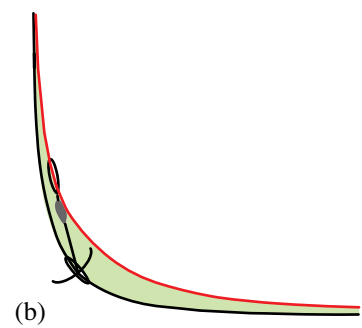
Here are two problems that can be easily solved by the method developed in this chapter. They are easy to understand but more challenging than they appear at first sight.

The reader may wish to try solving them before reading the chapter.

The tracks traced by the front and rear wheels of a bicycle of length 3 are shown below. In (a), the rear wheel follows an ellipse with cartesian equation $x^2 + 16y^2 = 16$, and in (b) the front wheel follows the hyperbola $y^2 - 3x^2 = 3$.



(a)



(b)

Find the area of the shaded region between the front and rear tracks in each case.

CONTENTS

Many standard problems in calculus can be easily solved by an innovative visual approach that makes no use of formulas. The method is based on Mamikon's sweeping-tangent theorem, a geometrically intuitive result that is easily understood by very young students. This chapter introduces the method of sweeping tangents and shows how it can be used to find (without the machinery of calculus) areas of many plane regions, including an oval ring, a parabolic segment, a hyperbolic segment, the region under a general power function, an exponential curve, a logarithmic curve, a tractrix, the region between two curves traced by the rear and front wheels of a bicycle, the region enclosed by a cardioid, and by each member of a family of limaçons. The treatment of the parabolic segment and the exponential use geometric properties of subtangents to these curves, which can also be used to draw tangent lines.

An unexpected application of the method of sweeping tangents is to physics. In this application, conservation of angular momentum in a central force field is deduced as an elementary consequence of Mamikon's sweeping-tangent theorem. Subsequent chapters give further applications of the method to both area and arclength, and also to 3-dimensional problems involving volume and surface area of solids.

1.1 INTRODUCTION

Calculus is a beautiful subject with a host of dazzling applications. Anyone familiar with this important branch of mathematics would be amazed to learn that many standard calculus problems can be easily solved by an innovative visual approach that makes no use of formulas. Here's a sample:

Problem 1. *Find the area of a parabolic segment.*

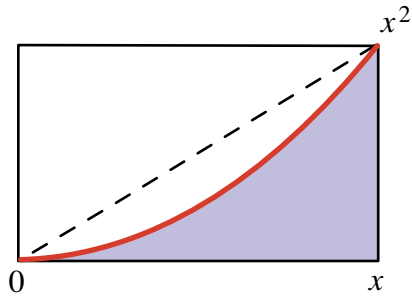


Figure 1.1: A parabolic segment.

In Figure 1.1 the parabolic segment is the shaded region between the graph of $y = x^2$ and the x axis from 0 to x. Its area was first calculated by Archimedes more than 2000 years ago by a method that laid the foundations for integral calculus. Today, every freshman calculus student can solve this problem: Integration of x^2 gives $x^3/3$. A solution without calculus is given in Section 1.10.

Problem 2. *Find the area of a region under an exponential curve.*

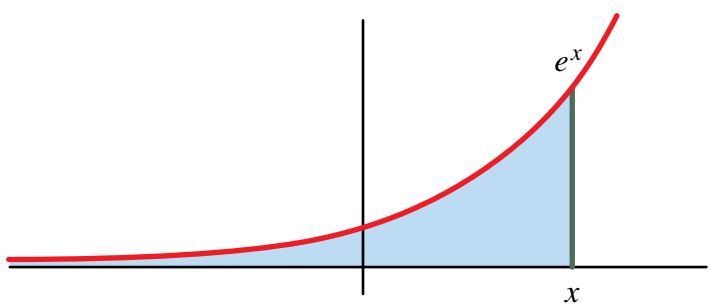


Figure 1.2: A region under an exponential curve.

Figure 1.2 shows the graph of the exponential function $y = e^x$, and we want the area of the shaded region between the curve and the x axis from minus infinity up to x. Integral calculus reveals that the answer is e^x. More generally, if the equation of the curve is $y = e^{x/b}$, where b is a positive constant, integration tells us that the area is $be^{x/b}$. A solution without calculus is given in Section 1.8.

Problem 3. *Find the area of the region under one arch of a cycloid.*

A cycloid is the path traced by a fixed point on the boundary of a circular disk that rolls along a line. For example, a light fastened to the rim of a bicycle wheel traces a cycloid as the bicycle rolls along a horizontal line. We want the area of the shaded region in Figure 1.3.

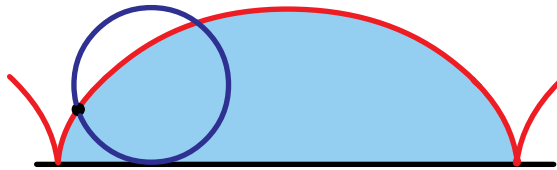


Figure 1.3: The region under one arch of a cycloid.

This classical calculus problem is more difficult than the first two. First, you need an equation for the cycloid, which takes some effort to derive. Integration shows that the area is exactly three times that of the rolling disk, a result we shall prove in Chapter 2 without using equations or calculus.

These three classical problems can be solved by a new method that relies on geometric intuition and is easily understood even by very young students. The new method does not require equations or integration. Moreover, it also solves some problems that cannot be done with calculus.

For example, look at the path traced by the front wheel of a moving bicycle as in Figure 1.4. The rear wheel traces another curve. What is the area of the region between the two tracks? To answer this question using calculus you need equations for the curves. But, with this new visual approach, no equations are needed and you can solve the problem easily, regardless of the shape of the bike's path. A solution is given in Section 1.4.

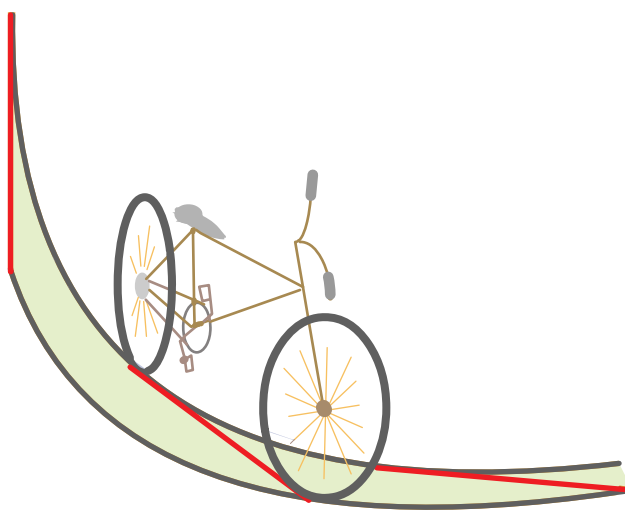


Figure 1.4: The region between the tire tracks of a bicycle.

1.2 EVOLUTION OF MAMIKON'S THEOREM

Like all great discoveries, Mamikon's method is based on a simple idea. It evolved half a century ago, when young Mamikon was presented with a classical geometry problem involving two concentric circles, with a chord of the outer circle tangent to the inner one, as in Figure 1.5.

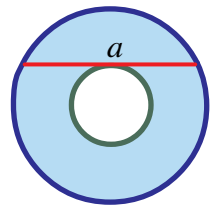


Figure 1.5: An annular ring between concentric circles.

The chord has length a, and the problem is: *Find the area of the annular ring between the circles.*

To solve it, look at Figure 1.6. The area of the inner circle of radius r is πr^2, and the area of the outer circle of radius R is πR^2, so the area of the ring is $\pi R^2 - \pi r^2 = \pi(R^2 - r^2)$. But the two radii and the tangent segment form a right triangle with legs r and $a/2$, and hypotenuse R, so by the Pythagorean Theorem, $R^2 - r^2 = (a/2)^2$, so the ring has area $\pi a^2/4$. The final answer depends only on a and not on the radii of the two circles!

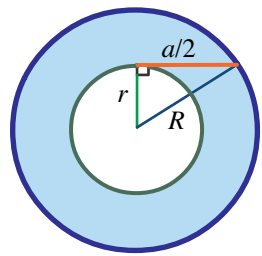


Figure 1.6: Ring with inner radius r and outer radius R.

If we knew in advance that the answer depends only on a, we could find it another way. Shrink the inner circle to a point, and the ring collapses to a disk of diameter a, with area equal to $\pi a^2/4$. An "Aha!" revelation, in the words of Martin Gardner.

Mamikon wondered if there was a way to see in advance why the answer depends only on the length of the chord. Then he thought of approaching the problem in a dynamic way. Take half the chord and think of it as a vector of length L tangent to the inner circle, as in Figure 1.7 (left). By moving the tangent vector around the inner circle, we see that it sweeps out the annular ring between the two circles, so the area is being swept by pure rotation.

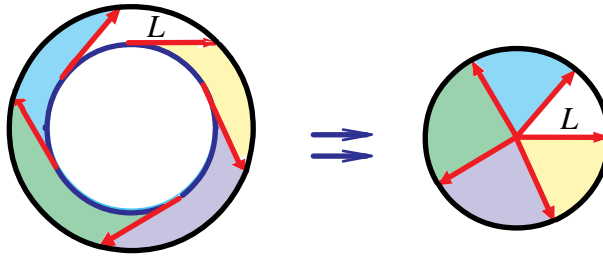


Figure 1.7: Ring swept out by a vector of constant length.

Now, translate each tangent vector parallel to itself so that the point of tangency is brought to a common point, as in Figure 1.7 (right). As the tangent vector moves around the inner circle, the translated vector rotates once around this common point and traces out a circular disk of radius L. So the tangent vectors sweep out a circular disk, as though they were all centered at the same point. And this disk has the same area as the ring.

Mamikon realized that this dynamic approach would also work if the inner circle is replaced by an arbitrary oval curve. Figure 1.8 shows the same idea applied to two ellipses. As a tangent segment of constant length moves once around each ellipse, it sweeps out a more general annular shape that we call an *oval ring*.

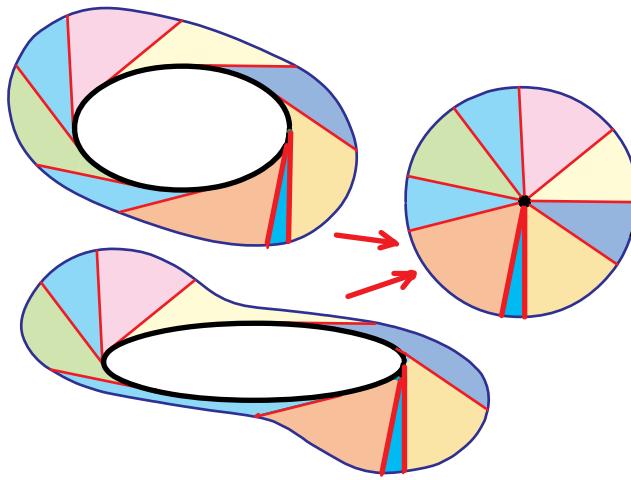


Figure 1.8: Tangent to an ellipse sweeps out an oval ring.

Again we translate each tangent segment parallel to itself so the point of tangency is brought to a common point. As the tangent moves around the oval, the translated segments trace out a circular disk whose radius is that constant length. So, the area of the oval ring should be the area of the circular disk.

The Pythagorean Theorem won't yield the areas of the oval rings. If the inner

oval is an ellipse you can calculate the areas by integral calculus (not a trivial task); if you do so, you'll find that all the oval rings have equal areas depending only on the length of the tangent segment!

Is it possible that the same is true for any convex simple closed curve? Figure 1.9 illustrates the idea for a triangle. As the tangent segment moves along an edge, it doesn't change direction so it doesn't sweep out any area.

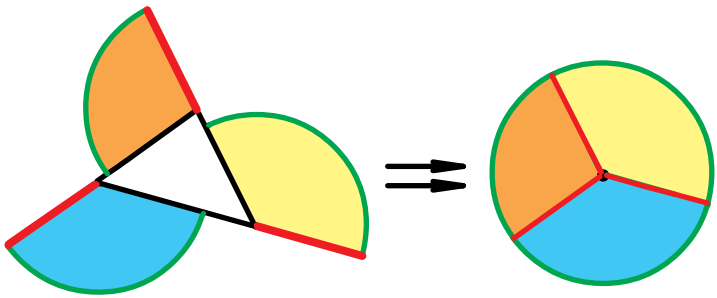


Figure 1.9: Tangent of constant length moving around a triangle.

As it moves around a vertex from one edge to the next, it sweeps out part of a circular sector. And as it goes around the entire triangle it sweeps out three circular sectors that, together, fill out a circular disk, as shown in Figure 1.9 (right). The same holds for any convex polygon, illustrated by a hexagon in Figure 1.10.

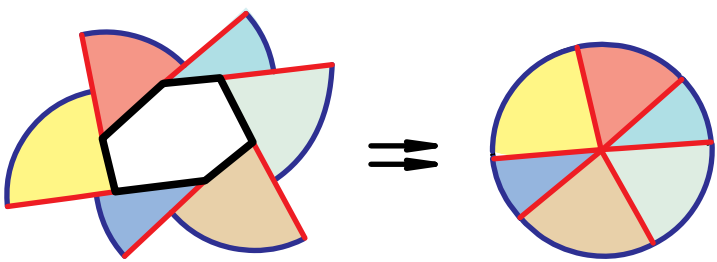


Figure 1.10: Constant-length tangent moving around a hexagon.

The area of the region swept out by a tangent segment of given length moving around a convex polygon is equal to the area of a circular disk whose radius is that length. Therefore the same is true for any convex curve that is a limit of convex polygons as the number of edges tends to infinity, as suggested by the example in Figure 1.11.

This leads us to:

Mamikon's Theorem for Oval Rings. *All oval rings swept by a line segment of given length, with one endpoint tangent to a smooth closed plane curve, have equal areas, regardless of the size or shape of the inner curve. Moreover, the area depends only on the length L of the tangent segment, and is equal to πL^2, the area of a disk of radius L, as if the tangent segment was rotated about its endpoint.*

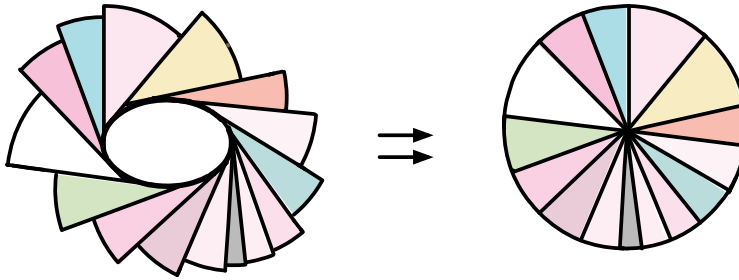


Figure 1.11: Constant-length tangent moving around a many-sided polygon.

Application to the Pythagorean Theorem.

Figure 1.12 illustrates how Mamikon's theorem can be used to provide a new proof of the Pythagorean Theorem. If the inner curve is a circle of radius r, the outer curve is also a circle (of radius R, say), so the area of the oval ring is equal to the

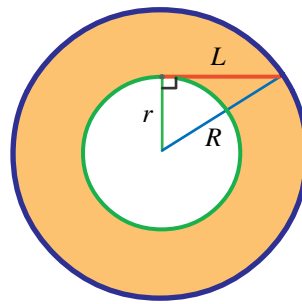


Figure 1.12: Mamikon's theorem implies the Pythagorean Theorem.

difference $\pi R^2 - \pi r^2$. But by Mamikon's theorem, the area of the oval ring is also equal to πL^2, where L is the constant length of the tangent segments. By equating areas we find $R^2 - r^2 = L^2$, from which we get $R^2 = r^2 + L^2$, the Pythagorean Theorem.

1.3 APPLICATION TO SLICES OF SPHERICAL SHELLS

A spherical shell is the region between two concentric solid spheres. Its cross section by a plane that intersects both the inner and outer spheres is an annular ring whose inner and outer radii depend on the cutting plane. Mamikon's theorem for circular rings implies a striking and somewhat unexpected result:

The area of an annular ring cut by a plane that intersects both spheres of a spherical shell is a constant independent of the position and inclination of the cutting plane.

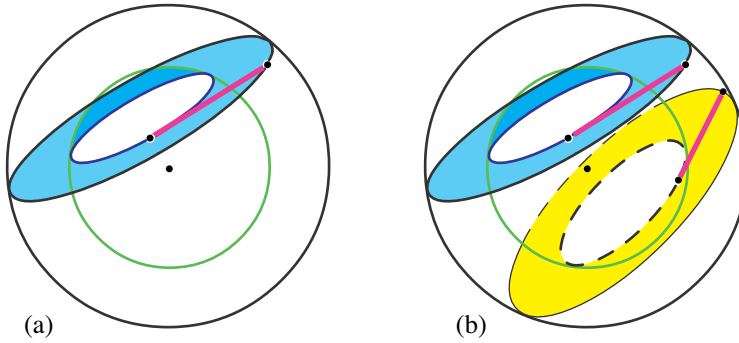


Figure 1.13: (a) A tangent to the inner circle is also tangent to the inner sphere. (b) Length of tangent segment is independent of position and inclination of the cutting plane.

Proof. From our earlier discussion we know that the area of an annular ring depends only on the length of the tangent segment to the inner circle cut off by the outer circle. It is easy to see that this length is the same for every plane that cuts both spheres, regardless of its position and inclination.

Simply observe that the tangent segment to the inner circle is also tangent to the inner sphere from a point on the surface of the outer sphere. Because of spherical symmetry, all such tangent segments to the inner sphere from the outer sphere have constant length (Figure 1.13a), for all positions and inclinations of the cutting plane, so long as it cuts both spheres (Figure 1.13b).

When the foregoing property is applied to slices of equal thickness cut by two parallel planes that intersect both spheres, we find that all slices have equal volume. This fact, in turn, implies that all the slices have equal outer surface area. It also has applications to tomography. Details appear in Chapters 5 and 15.

1.4 CONSTANT-LENGTH TANGENT SWEEP AND TANGENT CLUSTER OF A PLANE CURVE

Figure 1.14 shows a more general version of Mamikon's theorem. The lower curve on the left is a more or less arbitrary smooth plane curve that we call the *tangency curve* τ. The set of all tangent segments of constant length defines a region that is bounded by τ and an upper curve (called the *free-end curve* σ) traced by the tangent segment's other extremity. The exact shape of this region depends on curve τ and on the length of the tangent segments from τ to σ. We refer to this region as a *tangent sweep.*

When each tangent segment is translated parallel to itself to bring the points of tangency together as before, the set of translated segments is called a *tangent cluster.* Figure 1.14 shows a tangent sweep (left) and its corresponding tangent cluster (right).

Because the tangent segments have constant length, the tangent cluster in Figure 1.14 is a circular sector whose radius is that constant length. By the way, we could

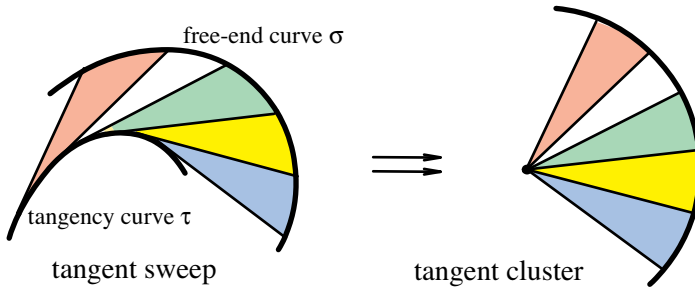


Figure 1.14: For a constant-length tangent sweep, the tangent cluster is a circular sector.

also translate the tangent segments so the other endpoints are brought to a common point. The resulting tangent cluster is a symmetric version of the other one. So we can state:

Mamikon's Theorem: Tangent Segments of Constant Length. *The area of a tangent sweep is equal to the area of its tangent cluster, regardless of the shape of the original curve.*

You can see this applied in a real-world illustration when a bicycle's front wheel traces out one curve, while the rear wheel (at constant distance from the front wheel) traces out another curve, as in Figure 1.15 (left).

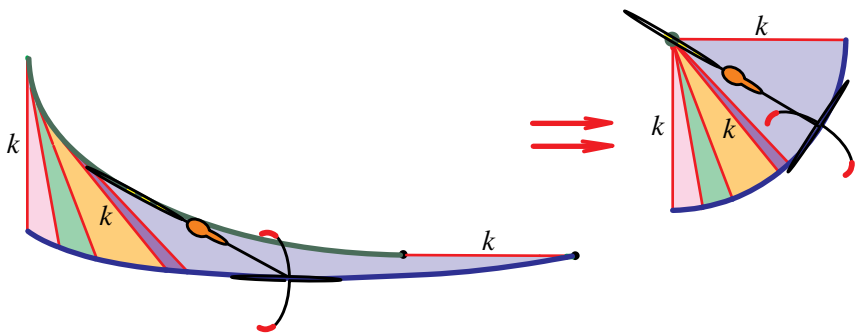


Figure 1.15: Finding the area of the region between two tire tracks.

The area of the tangent sweep is equal to the area of a circular sector depending only on the length of the bicycle and the change in angle from its initial position to its final position, as shown in the tangent cluster to the right in Figure 1.15. For more variations of the bicycle's path, see Section 15.6.

1.5 VARIABLE-LENGTH TANGENT SWEEP. SPACE CURVES

Figure 1.16 illustrates the same ideas in a more general setting. Now the tangent segments from tangency curve τ to free-end curve σ need not have constant length. We still have the tangent sweep (left) and the tangent cluster (right), formed by bringing the points of tangency together at a common point F.

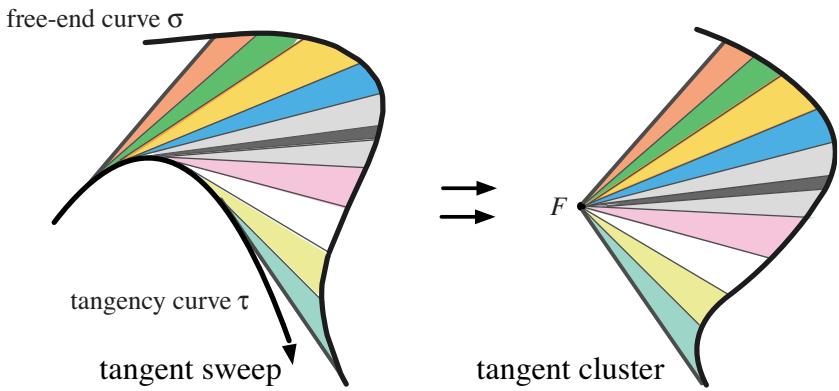


Figure 1.16: Variable-length tangent sweep and tangent cluster have equal areas.

Mamikon's theorem, which seems intuitively obvious by now, is that the area of the tangent cluster is equal to the area of the tangent sweep. To convince yourself, consider corresponding equal tiny triangles translated from the tangent sweep to the tangent cluster.

In the most general form of Mamikon's theorem, the tangency curve need not lie in a plane. It can be any smooth curve in space, and the tangent segments can vary in length, as in Figure 1.17.

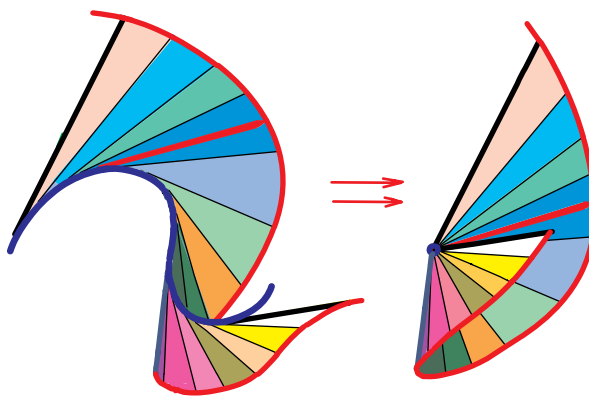


Figure 1.17: Variable-length tangent sweep and tangent cluster in space.

The tangent sweep will lie on a developable surface, one that can be rolled out

flat onto a plane without distortion. The shape of the tangent sweep depends on how the lengths and directions of the tangent segments change along the curve. The tangent cluster lies on a conical surface whose vertex is the common point. As expected, the area of the tangent sweep is equal to that of its tangent cluster.

General Form of Mamikon's Theorem. *The area of a tangent sweep to a space curve is equal to the area of its tangent cluster on a conical surface.*

This theorem, suggested by geometric intuition, is proved in Section 15.10 using differential geometry. Its importance stems from its wide variety of interesting and unexpected applications.

1.6 APPLICATION TO THE TRACTRIX

As already mentioned, curves swept by tangent segments of constant length include oval rings and bicycle-tire tracks. Another such example is the *tractrix*, the trajectory of a toy on a taut string being pulled by a child walking along a fixed straight line, as shown in Figure 1.18.

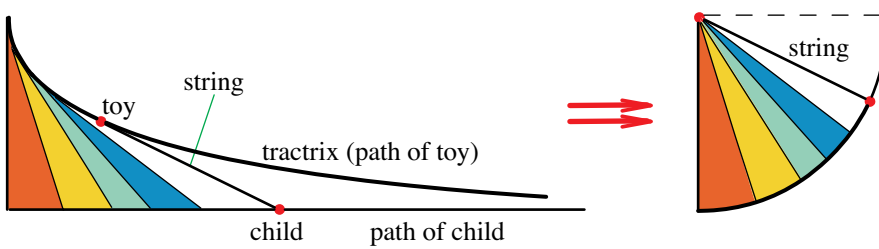


Figure 1.18: Tangent sweep and tangent cluster for a tractrix.

To find the area of the region between the tractrix and the x axis using calculus, you have to find the equation of the tractrix. This in itself is rather challenging because it requires solving a differential equation, as shown in Section 15.3. Once you have the equation of the tractrix you can use integration to find the area. This can also be done, but the calculation (given in Section 15.3) is somewhat demanding; the final answer is $\pi k^2/4$, where k is the length of the string. To find this without calculus, simply note that the tractrix is a particular case of the "bicyclix," whose tangent sweep is a circular sector of radius k. The full area is that of a quarter of a circular disk, or $\pi k^2/4$.

All these examples with tangents of constant length reveal the striking property that the area of the tangent sweep can be expressed in terms of the area of a circular sector without using any of the formal machinery of traditional calculus.

But even more striking are applications to examples in which the tangent segments are of variable length. These examples reveal the true power of Mamikon's method. Before we turn to them we digress briefly to discuss subtangents, which provide a simple geometric method of drawing tangent lines to curves.

1.7 SUBTANGENTS USED TO DRAW TANGENT LINES TO PLANE CURVES

The tangent line at a point $(x, f(x))$ on the graph of a function f is the line through that point with slope $f'(x)$. The simplest way to draw this line in practice, whether by hand or on a computer, is to join the point $(x, f(x))$ with another point known to be on the tangent line. Sometimes we can find such a point without explicitly calculating the slope $f'(x)$.

We illustrate with three exponential curves in Figure 1.19. In the first, the line

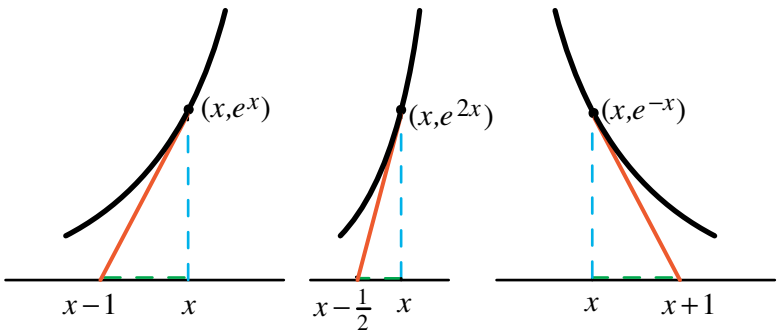


Figure 1.19: A simple way to draw tangents to exponential curves.

joining the point $(x-1, 0)$ with (x, e^x) is tangent to the graph of $f(x) = e^x$ because it has the required slope, $f'(x) = e^x$. In the second, the tangent joins $(x-\frac{1}{2}, 0)$ with (x, e^{2x}), and in the third, the tangent joins $(x+1, 0)$ with (x, e^{-x}).

On a general plane curve $y = f(x)$, the magic point on the x axis that also lies on the tangent line through $(x, f(x))$ is given by $(x - s(x), 0)$, where $s(x)$ is the subtangent defined by

$$s(x) = \frac{f(x)}{f'(x)} \tag{1.1}$$

at each point x where the derivative $f'(x)$ is nonzero. In Figure 1.20, $s(x)$ represents the base of a right triangle of altitude $f(x)$ and hypotenuse of slope $f'(x)$. From (1.1) we find $f'(x) = f(x)/s(x)$, so, if $s(x)$ is known, this gives a simple geometric procedure for finding the tangent line to an arbitrary point on the graph of f. As in Figure 1.20, drop a perpendicular from $(x, f(x))$ to point $(x, 0)$ on the x axis. Move to the point $(x - s(x), 0)$ on the x axis, and join it to $(x, f(x))$ to get the required tangent line.

The method of construction is especially useful when $s(x)$ has a simple form as in the following examples.

Example 1 (Constant subtangents). Exponential curves were first introduced in 1684 when Leibniz posed the problem of finding all curves with constant subtangents. The solutions are the exponential curves. Specifically, given a nonzero constant b we have $f(x) = Ke^{x/b}$ for some constant $K \neq 0$ if and only if $s(x) = b$.

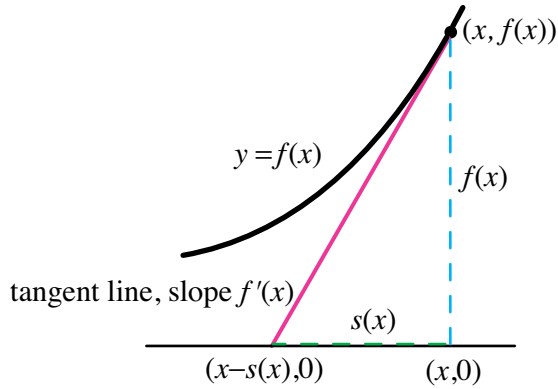


Figure 1.20: Geometric meaning of subtangent. The tangent line at $(x, f(x))$ passes through the point $(x - s(x), 0)$ on the x axis.

Examples with $b = 1, 1/2$, and -1 are shown in Figure 1.19. Incidentally, multiplying a function f by a nonzero constant does not alter its subtangent because f' is multiplied by the same factor, which cancels in (1.1).

In Section 1.8 we will use the constant subtangent property of exponential curves together with Mamikon's Theorem to show visually that the area of the region between the graph of $y = e^{x/b}$ and an arbitrary interval $(-\infty, x]$ is $be^{x/b}$, the same result one would get by integration.

Example 2 (Linear subtangents). Power functions have linear subtangents. In fact, for a nonzero constant b we have $f(x) = Kx^{1/b}$ for a constant $K \neq 0$ if and only if $s(x) = bx$. In particular, the parabola $f(x) = x^2$ has subtangent $s(x) = x/2$, and the hyperbola $f(x) = 1/x$ has subtangent $s(x) = -x$. Figure 1.21 (left) shows how tangent lines to the parabola $f(x) = x^2$ can be easily constructed by joining $(x/2, 0)$ to (x, x^2). Figure 1.21 (right) illustrates the tangent construction for the hyperbola $f(x) = 1/x$. Here $x - s(x) = 2x$, so the tangent line passes through the points $(2x, 0)$ and $(x, 1/x)$.

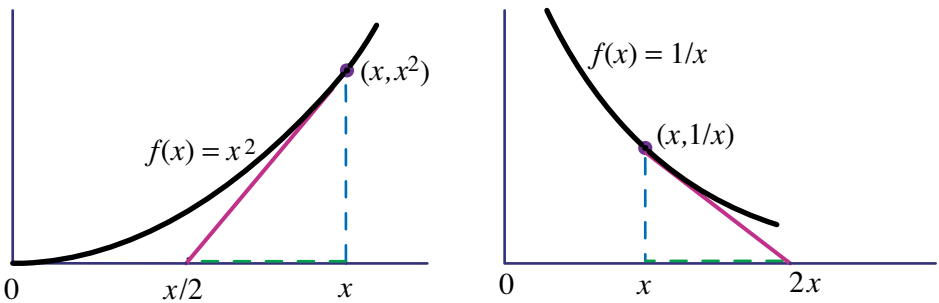


Figure 1.21: Simple geometric construction of tangents to the parabola $f(x) = x^2$, and the hyperbola $f(x) = 1/x$.

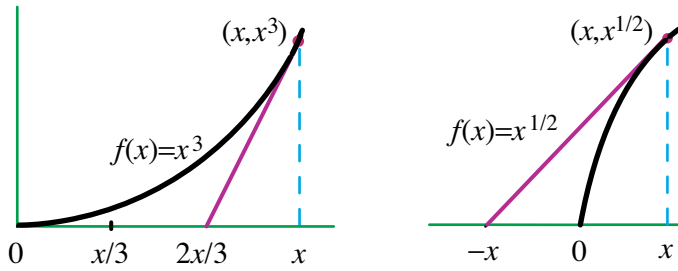


Figure 1.22: Simple geometric construction of tangents to $f(x) = x^3$, and to $f(x) = x^{1/2}$.

For the cubic curve $f(x) = x^3$ we have $s(x) = x/3$, so we join $(2x/3, 0)$ with (x, x^3) to get the tangent line at (x, x^3), as illustrated in Figure 1.22 (left). For the more general power function $f(x) = x^r$ we join $(x - x/r, 0)$ with (x, x^r) to get the tangent at (x, x^r). Figure 1.22 (right) furnishes an example with $r = 1/2$.

1.8 EXPONENTIAL CURVES

Exponential functions are ubiquitous in the applications of mathematics. They occur in problems concerning population growth, radioactive decay, heat flow, and other physical situations where the rate of growth of a quantity is proportional to the amount present. Geometrically, this means that the slope of the tangent line at each point of an exponential curve is proportional to the height of the curve at that point.

We have also seen that exponential curves are the only curves with constant subtangents. By exploiting this fact, we can use Mamikon's Theorem to find the area of the region under an exponential curve without using integral calculus. Figure 1.23 shows us how to do it.

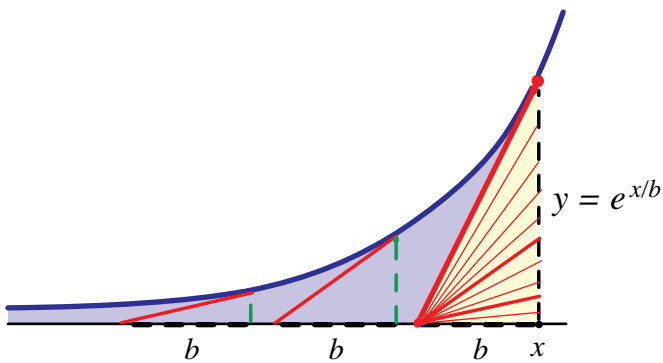


Figure 1.23: Region under an exponential curve swept by tangents.

The exponential curve $y = e^{x/b}$ has constant subtangents of length b, and the region in question is swept by the tangent segments cut off by the x axis as the

point of tangency moves left from x to minus infinity.

Figure 1.23 also shows each tangent segment translated so that the endpoint on the x axis is brought to a common point, in this case the lower vertex of a right triangle. Because the subtangent is constant, the resulting tangent cluster forms a right triangle of base b and altitude $e^{x/b}$. Therefore the area of the tangent sweep is equal to the area of the right triangle, so the area of the region between the exponential curve and the interval $(-\infty, x]$ is twice the area of the right triangle, which is its base b times its altitude $e^{x/b}$, or $be^{x/b}$.

In the language of calculus, we have shown that

$$\int_{-\infty}^{x} e^{t/b} dt = be^{x/b},$$

obtained as an application of Mamikon's Theorem, without the formal machinery of integral calculus.

1.9 AREA OF A HYPERBOLIC SEGMENT

The fact that exponential curves have constant subtangents can also be exploited to establish the classical formula

$$\int_{1}^{x} \frac{1}{t} dt = \log x, \tag{1.2}$$

where $\log x$ is the natural logarithm of x. If $x \geq 1$, the integral represents the area $A(x)$ of a hyperbolic segment, the region below the graph of the hyperbola $y = 1/x$ and above the interval $[1, x]$, shown in Figure 1.24.

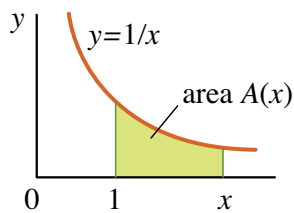


Figure 1.24: $A(x)$ is the area of the region below the hyperbola $y = 1/x$ and above the interval $[1, x]$.

In some calculus textbooks, (1.2) is taken as the definition of the logarithm function, and then the exponential function is defined to be its inverse function. (See [1], Secs. 6.3 and 6.12.) Here we adopt an alternative point of view. We define the exponential as that function with constant subtangent 1 and value 1 at 0, and define the logarithm as the inverse of the exponential. Now we show that the function $A(x)$ that describes the area of the hyperbolic segment in Figure 1.24, and which, in the language of calculus, is given by the integral

$$A(x) = \int_{1}^{x} \frac{1}{t} dt, \tag{1.3}$$

is the inverse of the exponential, which means $A(x) = \log x$, and gives (1.2).

Refer to Figure 1.25, which shows the general shape of the graph of the area function $y = A(x)$ as an increasing function of x for $x \geq 1$, with $A(1) = 0$. Differentiation of (1.3) gives $A'(x) = 1/x$, or $xA'(x) = 1$. Now we show geometrically that $xA'(x)$ represents the subtangent of the inverse of A, and because this subtangent is constant, the inverse is an exponential. Let B denote the inverse of A, so that

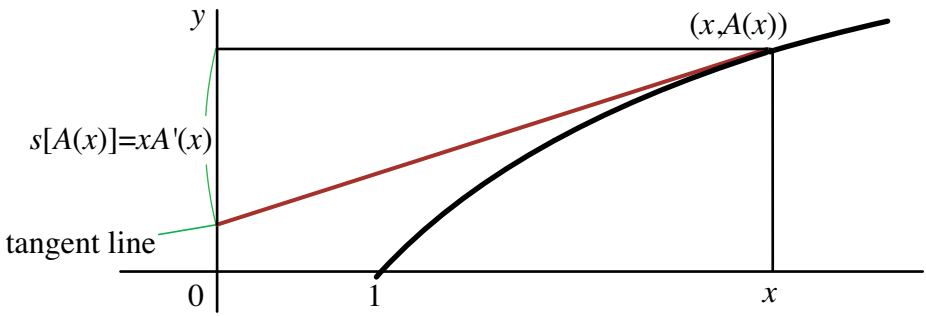


Figure 1.25: Tangent to $y = A(x)$ cuts off a segment on the y axis of length $xA'(x) = 1$.

$y = A(x)$ if, and only if, $x = B(y)$. Then $B[A(x)] = x$, which, when differentiated, gives $B'[A(x)]A'(x) = 1$. Hence

$$\frac{B[A(x)]}{B'[A(x)]} = \frac{x}{1/A'(x)} = xA'(x) = 1.$$

By (1.1), the subtangent function s associated with B is given by

$$s(y) = \frac{B(y)}{B'(y)},$$

and when $y = A(x)$, the foregoing equation tells us that $s(y) = 1$. In Figure 1.25, the tangent to the graph of $y = A(x)$ at $(x, A(x))$ cuts off a segment on the y axis of length 1, the subtangent of the inverse function B. Hence B is an exponential, $B(y) = e^y$, and its inverse is the logarithm, $A(x) = \log x$, which proves (1.2).

In Figure 1.25, the curve $y = \log x$ divides the rectangle of area xy into two regions. The upper region has area $e^y - 1 = x - 1$, so the lower region has area $xy - (x - 1)$. In the language of integral calculus, this states that $\int_1^x \log t\, dt = x \log x - x + 1$, a result derived here geometrically.

1.10 AREA OF A PARABOLIC SEGMENT

We turn next to Problem 1 in the Introduction, perhaps the oldest calculus problem in history−finding the area of a parabolic segment, shown shaded in Figure 1.26a.

The segment is inscribed in a rectangle of base x and altitude x^2. The area R of the rectangle is x^3, but we will not need this explicit formula for R. From Figure 1.26a it is clear that the area of the parabolic segment is less than $R/2$, half that

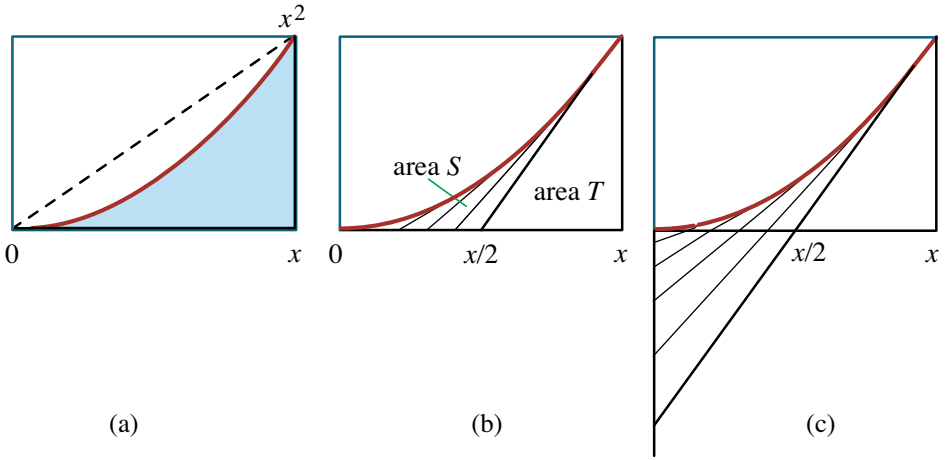


Figure 1.26: (a) Parabolic segment. (b) Tangent sweep of parabolic segment cut off by the x axis. (c) Region obtained by doubling the lengths of the tangent segments in (b).

of the rectangle. Archimedes made the stunning discovery that the area is exactly one-third that of the rectangle, or $R/3$. We will deduce this by a simple geometric approach using tangent sweeps.

The parabola has equation $y = x^2$, but we shall not need this. We use only the fact that the subtangent is $x/2$, so the tangent line above any point x cuts off a segment of length $x/2$, as in Figure 1.26b. The tangent sweep in Figure 1.26b is obtained by drawing all tangent lines to the parabola between 0 and x and cutting them off at the x axis.

We also use another property of tangents sweeps, called the scaling property, that is valid for both plane and space curves.

Scaling Property. *If each tangent segment of a tangent sweep is scaled* (*expanded or contracted*) *by the same positive factor t, then the area of the tangent sweep is multiplied by t^2.*

This follows because the tangent cluster of the scaled tangent sweep is also scaled radially by factor t producing a similar figure with area multiplied by t^2.

In Figure 1.26b the parabolic segment is divided into two regions, the tangent sweep, whose area we call S, and a right triangle, whose area we call T. We will prove that $S = T/3$, so the area of the parabolic segment, $S + T$, is equal to $4T/3$. Because $4T = R$, the area $S + T$ of the parabolic segment is $R/3$, as asserted.

To prove that $S = T/3$, refer to Figure 1.26c, where each tangent segment from Figure 1.26b has been doubled in length to reach the y axis, as shown. The area of the scaled tangent sweep is $4S$. But the expanded region consists of two parts, the portion above the x axis with area S, and the right triangle with area T below the x axis. Hence $4S = S + T$, so $S = T/3$. This proves that the parabolic segment has area $R/3$.

In the language of calculus, this simple argument shows that

$$\int_0^x t^2\,dt = \frac{x^3}{3},$$

a result called the quadrature of the parabola. It was first obtained by Archimedes using a geometric limiting process called the method of exhaustion. Interest in the method of exhaustion was revived in the 16th century, and the method was gradually transformed into a powerful discipline known as integral calculus, in which the quadrature of the parabola is a routine exercise.

1.11 REAL POSITIVE POWERS

Now replace x^2 by any power x^r with $r > 0$. When $f(x) = x^r$ the subtangent is x/r, so the tangent above any point x cuts off a segment on the x axis of length x/r. The example in Figure 1.27 has $r = 3$, and the tangent above x cuts off a segment of length $x/3$. We will use tangent sweeps to show that the cubic segment below the curve $y = x^3$ and above the interval $[0, x]$ has area $R/4$, one-fourth that of the circumscribing rectangle.

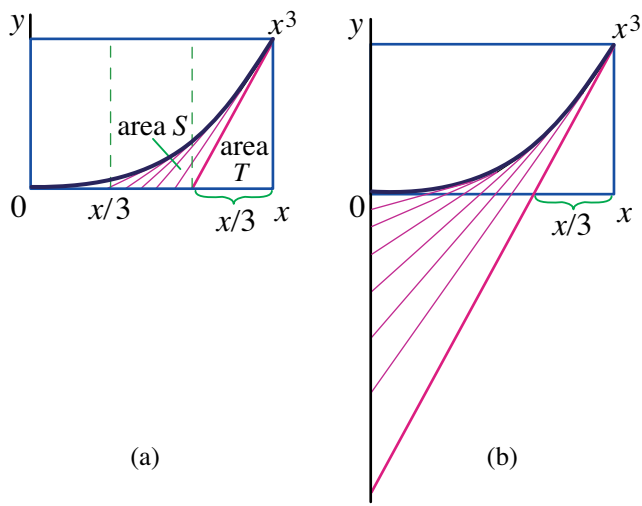


Figure 1.27: (a) Cubic segment divided into two regions. (b) Tangent segments in (a) tripled in length.

In Figure 1.27a the cubic segment is divided into two regions, the tangent sweep (of area S) and a right triangle (of area T). We will prove that $S = T/2$.

Expand each tangent segment in Figure 1.27a by a factor 3 so it reaches the y axis as shown in Figure 1.27b. The enlarged tangent sweep in Figure 1.27b has area $9S$. The portion of the enlarged region above the x axis has area S and the portion below the x axis is a right triangle with edges twice those in Figure 1.27a, so its area is $4T$. Therefore $9S = S + 4T$, so $S = T/2$. The area of the cubic segment is

$S + T = 3T/2 = 6T/4 = R/4$ because $6T$ is also the area R of the circumscribing rectangle. This shows that the area of the cubic segment is $R/4$, as claimed.

For a general power $r > 1$ the argument is similar. The general segment can be divided into two regions, a tangent sweep of area S cut off by the x axis, and a right triangle of area T. We expand each tangent segment by a factor r to obtain an enlarged tangent sweep of area r^2S cut off by the y axis. The portion of the enlarged region above the x axis has area S, and the portion below the x axis is a right triangle of area $(r-1)^2T$. Therefore $r^2S = S + (r-1)^2T$, so that $S = (r-1)^2T/(r^2-1) = (r-1)T/(r+1)$, and the area of the general segment is $S + T = 2rT/(r+1)$. But $2rT = R$, the area of the circumscribing rectangle, so the area of the general segment is $R/(r+1)$. In the language of integral calculus this is equivalent to

$$\int_0^x t^r dt = \frac{x^{r+1}}{r+1}.$$

The method also works if $r = 1$, giving $x^2/2$. In this case, the segment in question is a right triangle with area half that of the circumscribing rectangle. (The tangent sweep has area $S = 0$.)

If $0 < r < 1$ the graph changes shape from convex to concave, and the tangent sweep lies above the curve rather than below it. The analysis can be modified to cover this case, but the problem for $r < 1$ is easily reduced to that for exponents greater than 1. The example with $r = 1/2$ in Figure 1.28 shows us how to proceed in general.

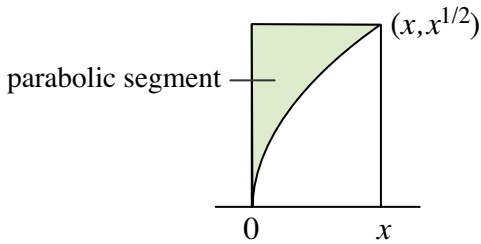


Figure 1.28: The area for $r = 1/2$ can be derived from that of the parabolic segment.

The portion of the rectangle above the graph is congruent to a parabolic segment with area $R/3$, so the portion below the graph has area $2R/3 = R/(1 + 1/2)$. (In calculus, this method amounts to integration by parts.)

For the general case $0 < r < 1$ the portion of the rectangle above the graph of $y = x^r$ is congruent to a general segment with exponent $1/r \geq 1$, so its area is equal to $R/(1 + 1/r)$. Therefore the portion below the graph has area $R - R/(1 + 1/r)$ $= R/(r + 1)$, the same formula obtained for $r > 1$.

1.12 GENERAL NEGATIVE POWERS

We turn next to the graph of $y = x^{-r}$, where $r > 1$, and we determine the area of the region over the interval $[x, \infty]$ for any $x > 0$. Integral calculus tells us that the area is given by

$$\int_x^\infty t^{-r}\, dt = \frac{x^{1-r}}{r-1}. \tag{1.4}$$

We will prove this geometrically without using integral calculus.

The general segment in question consists of two parts shown in Figure 1.29a, a right triangle of area $T = x^{1-r}/(2r)$ and a tangent sweep of area S, hence the segment has area $S+T$. This region is adjacent to a rectangle of base x and altitude x^{-r} whose area $R = x^{1-r} = 2rT$. We will show that $S + T = R/(r-1)$, the same result given by (1.4). Incidentally, this shows that S cannot be infinite.

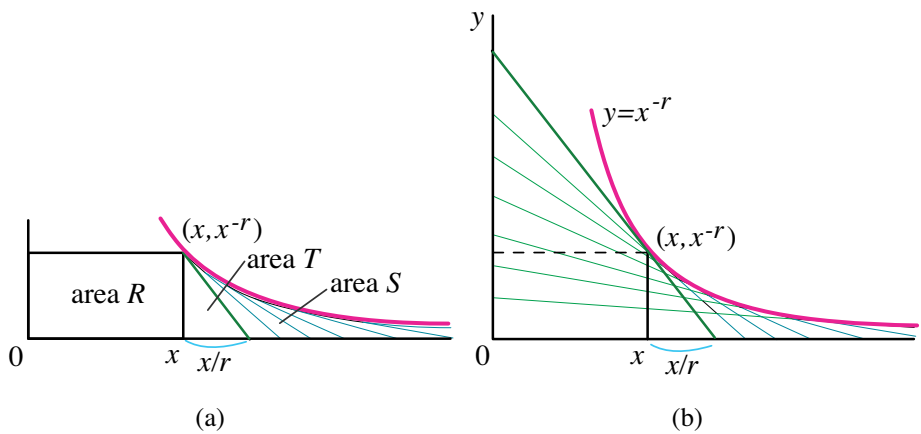


Figure 1.29: (a) A general segment divided into a tangent sweep of area S and a triangle of area T. (b) The tangent sweep cut off by the y axis has area r^2S.

Figure 1.29b shows the region obtained by drawing the tangent segments from each point of tangency (t, t^{-r}) to the y axis for all $t \geq x$. The length of each tangent segment is r times that of the tangent segment at the same point cut off by the x axis in Figure 1.29a, so the tangent sweep in Figure 1.29b has area r^2S. It consists of two parts, a right triangle of area $(r+1)^2T$, and the original tangent sweep of area S adjacent to it. Therefore $r^2S = (r+1)^2T + S$, which gives $S = (r+1)^2T/(r^2-1) = (r+1)T/(r-1)$. Hence

$$S + T = (\frac{r+1}{r-1} + 1)T = \frac{2rT}{r-1} = \frac{R}{r-1},$$

as required.

1.13 AN ALTERNATIVE APPROACH FOR NEGATIVE POWERS

Another geometric approach is illustrated in Figure 1.30, which shows a tangent cluster of the tangent sweep in Figure 1.29a obtained by translating each tangent segment parallel to itself so that the point of tangency is brought to the origin.

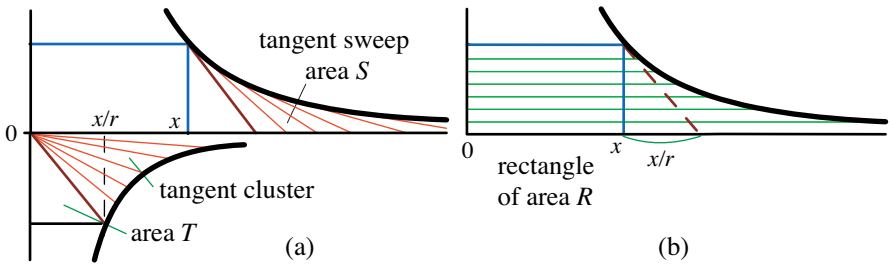


Figure 1.30: (a)Tangent cluster of the tangent sweep in Figure 1.29a adjacent to a triangle of area T. (b) Reflection of the lower region in (a) and horizontal stretching by r.

By Mamikon's Theorem, this tangent cluster has area S. The tangent cluster and the right triangle adjacent to it in Figure 1.30a have area $S+T$. Next, reflect this region (tangent cluster plus triangle) through the x axis, giving a congruent region of area $S+T$, then stretch it horizontally by a factor r (that is, multiply the x coordinate of each point in the reflected region by r) to get a new region of area $r(S+T)$ indicated by the horizonal shading in Figure 1.30b. The stretched region is made up of two parts, a rectangle of area R, and the original region with area $S+T$ in Figure 1.30a. Therefore $r(S+T) = R + (S+T)$, and again we find

$$S + T = \frac{R}{r-1}.$$

If $0 < r < 1$ the integral in (1.4) diverges, but in this case the area of the region under the curve $y = x^{-r}$ and above the interval $[0, x]$ is

$$\int_0^x t^{-r}\,dt = \frac{x^{1-r}}{1-r}.$$

This approach can also be adapted to positive powers.

1.14 A REVERSE TYPE OF APPLICATION

In each of the foregoing examples we found the area of a tangent sweep by equating it to the area of its corresponding tangent cluster, which was easier to determine. But if the area of the tangent sweep is easier to determine, then Mamikon's Theorem can be used as a two-edged sword to give the area of the tangent cluster. This section exploits this idea to give a geometric proof of the integration formula

$$\int_0^x \tan^2\,\theta\,d\theta = \tan x - x. \tag{1.5}$$

Figure 1.31 (left) shows the graph of the polar equation $r(\theta) = \tan\theta$ as θ varies from 0 to x. The area $A(x)$ of the shaded region is given by

$$A(x) = \frac{1}{2}\int_0^x \tan^2\theta\, d\theta.$$

Now we show geometrically that $A(x) = \frac{1}{2}\tan x - \frac{1}{2}x$.

The region in Figure 1.31 (left) is the tangent cluster of the tangent sweep to the unit circle, shown dashed on the right. Here each tangent segment to the unit circle is cut off by the vertical line through the center of the circle.

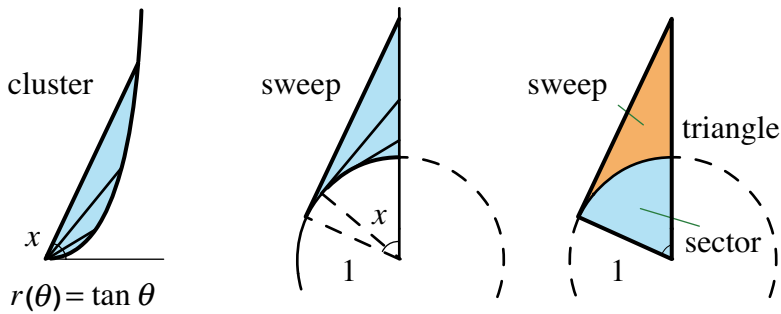


Figure 1.31: The shaded region on the left is the tangent cluster of the tangent sweep on the right. They have equal areas. The area of the tangent sweep is the area of a right triangle minus the area of a circular sector.

If the radius makes angle θ with the vertical line, then the corresponding tangent segment has length $\tan\theta$. The area $A(x)$ of the tangent sweep is equal to that of a right triangle of base 1 and altitude $\tan x$, which is $\frac{1}{2}\tan x$, minus the area of the circular sector subtending the angle x, which is $\frac{1}{2}x$. Therefore $A(x) = \frac{1}{2}\tan x - \frac{1}{2}x$ which, when multiplied by 2, gives (1.5).

1.15 APPLICATION TO THE LIMAÇON OF PASCAL

Limaçon as a pedal curve.

Start with a smooth curve Γ and a point P (which need not be on Γ) called a *pedal point*. Let F denote the foot of the perpendicular from P to an arbitrary tangent line to Γ. The locus of all such points F constructed for all tangent lines is called the pedal curve of Γ with respect to P. When Γ is a circle, as in Figure 1.32a, the pedal curve is called a *limaçon of Pascal*. Other pedal curves, with Γ not a circle, will occur in Chapter 3. In Section 3.13 the limaçon is also described as a roulette.

To show that the pedal description is well suited to the use of Mamikon's sweeping-tangent theorem, we calculate the area of the region enclosed by a limaçon, starting with the case where pedal point P is on Γ.

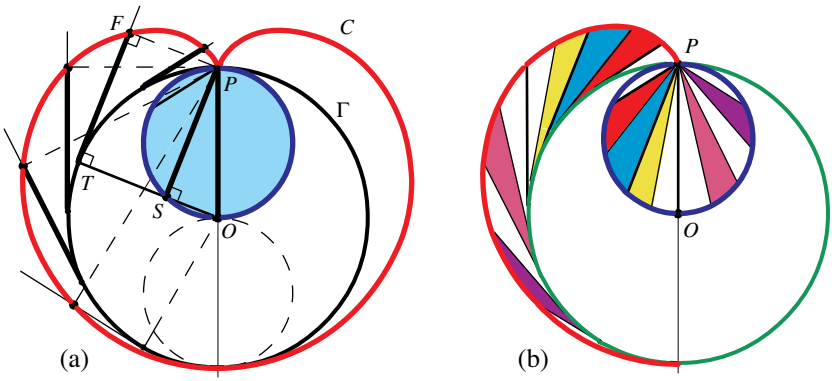


Figure 1.32: (a) Cardioid as a pedal curve with the pedal point P on circle Γ. (b) Tangent sweep of a lune and its tangent cluster, a circular disk. The lune and disk have equal areas.

Area of the region enclosed by a cardioid.

If pedal point P is on the circle, as in Figure 1.32a, the limaçon is a cardioid, denoted here by C. (A different description of a cardioid as an epicycloid will be given in Chapter 2, Section 2.4.) The cardioid C in Figure 1.32a encloses the circle Γ and two congruent crescent-shaped lunes between Γ and C. Each lune is a tangent sweep from Γ to C. A typical tangent segment from T on Γ to F on C is equal in length and parallel to a chord PS of the smaller circle with diameter OP, because $PFTS$ is a rectangle (angle PSO, inscribed in a semicircle, is a right angle). Therefore, the tangent cluster of each lune is a circular disk D of diameter OP (Figure 1.32b) whose area we denote by $[D]$. Thus, each lune has area $[D]$. The disk bounded by Γ has diameter $2OP$, so its area is $4[D]$. Hence the region bounded by the cardioid C has area $6[D]$.

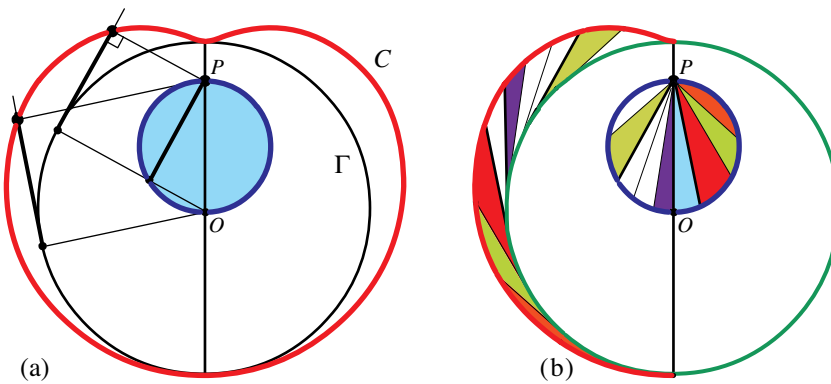


Figure 1.33: (a) Pedal curve C of circle Γ with respect to an interior pedal point P. (b) Tangent sweep of one lune and its tangent cluster, a circular disk.

Area of a region enclosed by a general limaçon.

Start with a circle Γ as in Figure 1.32, but choose the pedal point P inside Γ as shown in Figure 1.33a. This produces a family of limaçons C depending on the distance OP. Again, each lune is a tangent sweep from Γ to C, and its tangent cluster is the circular disk of diameter OP, as in Figure 1.33b. Thus the area of each lune is equal to the area $[D]$ of the disk D with diameter OP. The area of the region bounded by C is $2[D]$ plus the area of the Γ disk. If Γ has diameter d, then its area $[\Gamma]$ is $\lambda^2[D]$, where λ is the ratio of diameters, $\lambda = d/OP$.

If P is outside Γ as shown in Figure 1.34, the two lunes overlap, creating a loop as indicated in Figure 1.34a. Surprisingly, the area of each lune is $[D]$, the area of the disk of diameter OP, just as when P was inside or on Γ.

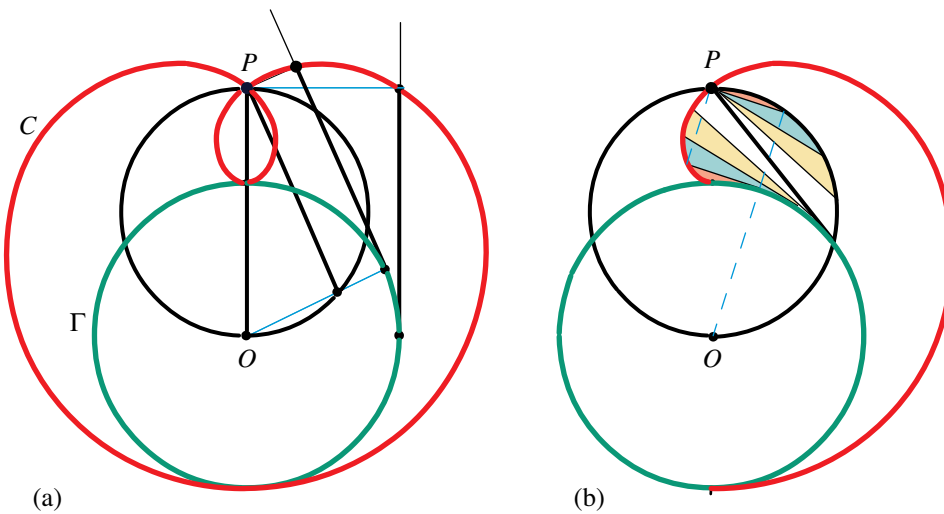


Figure 1.34: (a) Pedal curve C of Γ with respect to an exterior point P, with a loop through P. (b) The tangent sweep from Γ to C of the overlapping part of one lune, with its tangent cluster as a segment of the disk of diameter OP.

To see why, consider first Figure 1.34b, which shows the tangent sweep from Γ to C on the portion of one lune contributing to the overlapping loop. The corresponding tangent cluster is a segment of the disk of diameter OP, above the chord from Γ to P. This chord is also tangent to Γ.

Next consider Figure 1.35a, which shows part of the tangent sweep (from Γ to C) of the portion of the lune without the overlapping part, starting from the tangent through P and ending with the tangent parallel to OP. The corresponding tangent cluster is a segment of the disk of diameter OP above the chord from Γ to P.

Finally, Figure 1.35b shows that the remaining tangent sweep of the lune from Γ to C has as its tangent cluster the left half of the disk of diameter OP. Consequently, the area of one lune is equal to $[D]$, as asserted.

Thus, for all three cases, P on, inside, or outside circle Γ, we have:

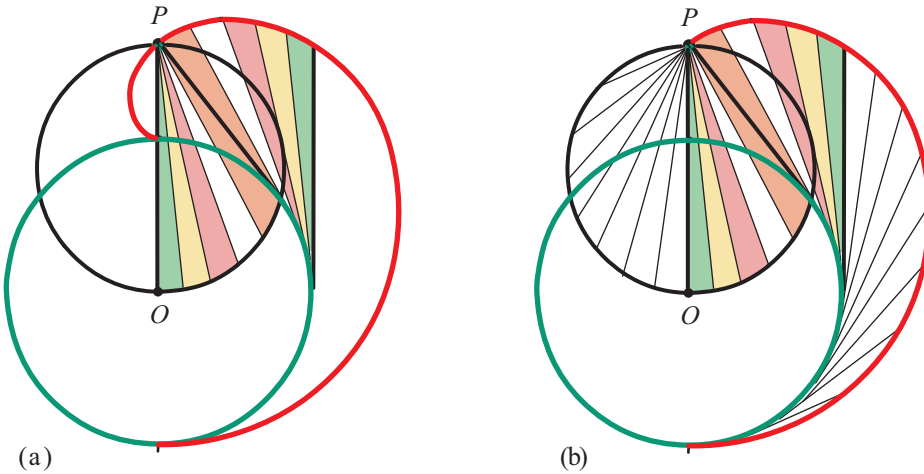


Figure 1.35: (a) Tangent sweep from Γ to C starting from the tangent through P and ending with the tangent parallel to OP, together with its tangent cluster. (b) The remaining tangent sweep from Γ to C has tangent cluster half the disk of diameter OP.

Each lune has area equal to the area $[D]$ *of the disk with diameter* OP.

To the best of our knowledge, this result has not been previously recorded.

When P is outside Γ, the area of the region enclosed by the entire limaçon C is the sum of the areas of the two overlapping lunes, which is $2[D]$, minus the area $[L]$ of the region L enclosed by the loop, plus the area $[\Gamma]$ of the Γ disk. To determine the loop area $[L]$ using sweeping tangents, refer to Figure 1.35b and note that, by symmetry, $[L]$ is twice the area of the left portion of the tangent sweep from Γ to C cut off by diameter OP. The right portion of the tangent sweep is exactly the same type that occurs in the tangent sweep in Figure 1.31, and its area is that of a right triangle of hypotenuse OP minus the area of a circular sector of the Γ disk.

Animated versions of the sweeping process for each of the three cases, P on, inside, and outside the circle Γ, are given on the web site

www.its.caltech.edu/~mamikon/calculus.html

For P on the circle, click on "CardioSwp", for P inside, click on "PodInArea", and for P outside, click on "PoderOutSwp".

1.16 APPLICATION TO PHYSICS

A fundamental law of physics states that angular momentum is conserved in a central force field. We describe this mathematically and show that it is a consequence of Mamikon's sweeping-tangent theorem.

Physics prerequisites.

Suppose the position of a moving particle is given by a radius vector $\boldsymbol{r}$ emanating from a fixed point $\boldsymbol{O}$. We regard $\boldsymbol{r}$ as a vector-valued function of time t. The free end of $\boldsymbol{r}$ traces a path along which the motion takes place, and the length $|\boldsymbol{r}|$ represents the distance of the particle from $\boldsymbol{O}$ at time t. The velocity $\boldsymbol{\nu}$ of the particle is defined as $\boldsymbol{\nu} = d\boldsymbol{r}/dt$, the time-derivative of $\boldsymbol{r}$, and the acceleration vector $\boldsymbol{a}$ is defined as $\boldsymbol{a} = d\boldsymbol{\nu}/dt$, the time-derivative of $\boldsymbol{\nu}$. The velocity vector is always tangent to the path of the particle and points in the direction of motion. Its length $|\boldsymbol{\nu}|$ represents the speed of the particle. The path traced by the free end of $\boldsymbol{r}$ is called the tangency curve and is denoted by τ. If the particle has mass m, the vector $m\boldsymbol{\nu}$ is called the *momentum* of the particle, and the cross product $\boldsymbol{r} \times m\boldsymbol{\nu}$ is called its *angular momentum*.

The cross product $\boldsymbol{r} \times \boldsymbol{\nu}$ of two vectors $\boldsymbol{r}$ and $\boldsymbol{\nu}$ is perpendicular to the plane of $\boldsymbol{r}$ and $\boldsymbol{\nu}$, and its length is $|\boldsymbol{r}||\boldsymbol{\nu}|$ times the sine of the angle between $\boldsymbol{r}$ and $\boldsymbol{\nu}$. In particular, the cross product of two parallel vectors is the zero vector $\boldsymbol{0}$. When the vectors $\boldsymbol{r}$ and $\boldsymbol{\nu}$ are placed as indicated in Figure 1.36, they form two edges of a parallelogram which, of course, may change its size and position during the motion.

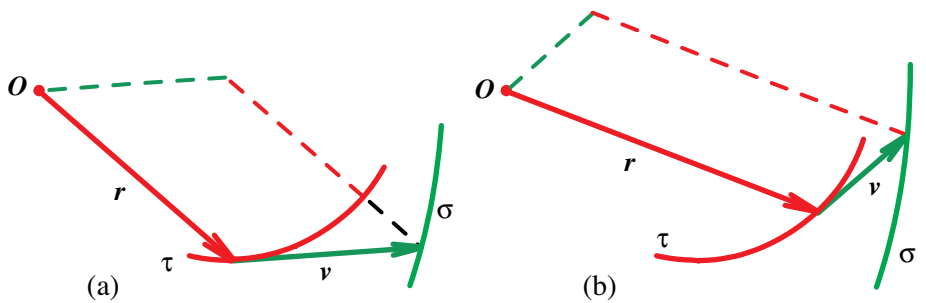


Figure 1.36: Parallelogram formed by vectors $\boldsymbol{r}$ and $\boldsymbol{\nu}$ at two different times.

The free end of $\boldsymbol{\nu}$ traces a curve we have denoted by σ. At time t the parallelogram has opposite vertices at $\boldsymbol{O}$ and on the curve σ, and its area $A(t)$ is given by

$$A(t) = |\boldsymbol{r} \times \boldsymbol{\nu}|. \tag{1.6}$$

The scalar quantity $mA(t)$ is the length of the angular momentum vector at time t.

Motion with radial acceleration.

When the acceleration vector $\boldsymbol{a}$ is always parallel to the radius vector $\boldsymbol{r}$, the motion is said to have *radial acceleration*. If the motion is caused by a force $\boldsymbol{F}$ acting on the particle, Newton's second law of motion states that $\boldsymbol{F} = m\boldsymbol{a}$, hence the acceleration vector is always parallel to the force vector. A *central force field* is one with $\boldsymbol{F}$ always parallel to the radius vector, and hence it produces radial acceleration.

It is easy to show that in a central force field, the parallelogram area $A(t)$ in (1.6) is constant (independent of t). This is because the derivative of the cross

product vector is the zero vector $\mathbf{0}$. In fact,

$$\frac{d}{dt}(\boldsymbol{r}\times\boldsymbol{\nu}) = \boldsymbol{r}\times\frac{d\boldsymbol{\nu}}{dt} + \frac{d\boldsymbol{r}}{dt}\times\boldsymbol{\nu} = \boldsymbol{r}\times\boldsymbol{a} + \boldsymbol{\nu}\times\boldsymbol{\nu} = \mathbf{0}$$

because each cross product $\boldsymbol{r}\times\boldsymbol{a}$ and $\boldsymbol{\nu}\times\boldsymbol{\nu}$ of parallel vectors is zero. Therefore $\boldsymbol{r}\times\boldsymbol{\nu}$ is a constant vector, so its length $A(t)$ is also constant.

Geometric derivation of the law of conservation of angular momentum in a central force field.

Figure 1.37a shows the tangent sweep traced by the velocity vector $\boldsymbol{\nu}$, and also its tangent cluster, obtained by translating each tangency point to $\boldsymbol{O}$.

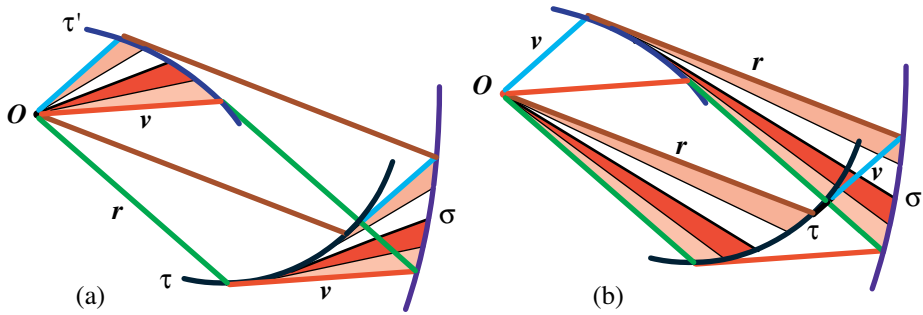


Figure 1.37: (a) The tangent sweep of the velocity vector has the same area as its tangent cluster. (b) For a central force field, the tangent sweep of tangent lines to the hodograph cut off by curve σ has the same area as its tangent cluster, which is swept by the radius vector from $\boldsymbol{O}$ to τ.

We know by Mamikon's Theorem that the tangent sweep and tangent cluster have equal areas. The free end of the translated velocity vector traces a curve denoted in Figure 1.37a by τ', called the *hodograph* by Hamilton [44], who used it to deduce the law of gravitation from the fact that planets move in elliptical orbits about the Sun. Hamilton made no use of the area of the tangent cluster and missed an opportunity to deduce the law of conservation of angular momentum in a central force field by a simple geometric argument, to which we turn next.

Now we will show that $A(t)$ is constant by using Mamikon's sweeping tangent theorem twice, following a proof communicated to the authors by Lang Withers [69]. Figure 1.37b shows the tangent sweep of the tangent lines to the hodograph τ' cut off by σ. Because the acceleration is radial, when the tangent segments to τ' are translated so each point of tangency is moved to $\boldsymbol{O}$, the corresponding tangent cluster is the region swept by the radius vector from $\boldsymbol{O}$ to τ, and it has the same area as the tangent sweep from the hodograph to σ.

From this it is easy to show that the area of the parallelogram generated by $\boldsymbol{r}$ and $\boldsymbol{\nu}$ remains unchanged during the motion. Rearrange the shaded regions in Figure 1.37 as shown in Figure 1.38. The sum of the areas of the two shaded regions

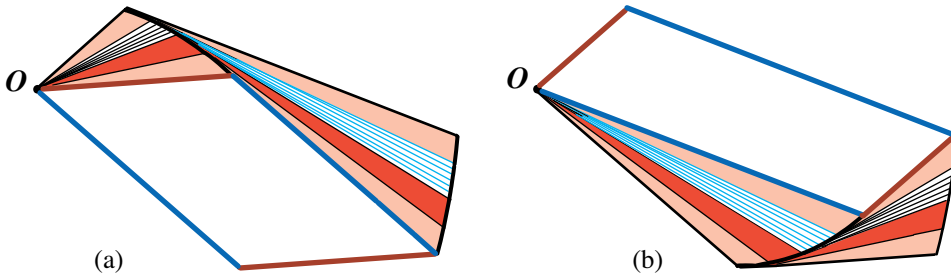


Figure 1.38: (a) The area of the entire figure is equal to the area of a parallelogram plus the sum of the areas of the two shaded regions. (b) The area of the same figure is equal to the area of another parallelogram plus the same sum of the areas of the two shaded regions, so the two parallelograms have equal areas.

in Figure 1.38a is equal to the sum of the areas of the two shaded regions in Figure 1.38b. From this it is obvious that the parallelograms in Figure 1.38a and 1.38b have equal areas.

NOTES ON CHAPTER 1

Much of this chapter is adapted from a lecture [3] delivered by Tom M. Apostol on October 4, 2000 at a colloquium held in honor of his first 50 years at the California Institute of Technology. The use of subtangents in connection with Mamikon's method was introduced in [8] and [10].

The chapter displays a wide canvas of topics that can be treated with Mamikon's method. To see them in animated form, consult the web site:

http://www.its.caltech.edu/~mamikon/calculus.html

Animation reveals in a dynamic way how tangent sweeps are generated, and how the tangent segments can be translated to form tangent clusters. Animation also reveals that many classical curves can be generated in a natural way by their intrinsic geometric and mechanical properties.

In subsequent chapters we apply Mamikon's method to many plane curves not treated above, for example, cycloids, epicycloids, hypocycloids, spirals, and pursuit curves. Applications to arclength are given in Chapters 3 and 11.

The method can also be used to find volumes of solids and their surface areas. Chapter 5 treats spheres and spherical shells, and generalizations called Archimedean domes and shells. Chapter 15 gives applications to tomography.

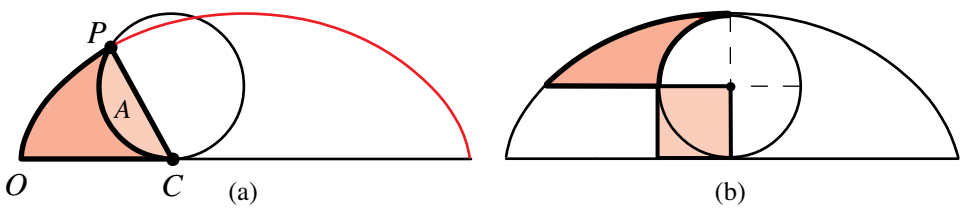
Finally, a philosophical remark about calculus. Newton and Leibniz, the discoverers of integral calculus, unified work done by many other pioneers, and related the processes of differentiation and integration. A connection between differentiation and integration is also imbedded in Mamikon's method, because it relates moving tangent segments with the areas of the swept regions.

Chapter 2

CYCLOIDS AND TROCHOIDS

The following problems can be easily solved by the methods developed in this chapter. The reader may wish to try solving them before reading the chapter.

When a circular disk rolls along a horizontal line, a point P on its circumference traces a cycloid, shown in (a) and (b). In (a), the tracing point has moved from O to P while the disk has rolled along the horizontal from O to C. The area A of the circular segment cut off by chord PC is known.



Show that the area of the tusk-shaped region OPC bounded by cycloidal arc OP, circular arc PC, and the horizontal segment OC, is $2A$.

Refer to (b). *Show that the shaded curvilinear region has the same area as the adjacent square.*

CONTENTS

A point on the boundary of a circular disk that rolls once along a straight line traces a cycloid. The cycloid divides its circumscribing rectangle into a cycloidal arch below the curve and a cycloidal cap above it. The area of the arch is three times that of the disk, and the area of the cap is equal to that of the disk.

This chapter provides deeper insight into this well-known property by applying Mamikon's sweeping-tangent theorem to show that the ratio 3:1 holds at every stage of rotation. Each cycloidal sector swept by the normal segment from the point of contact of the disk to the cycloid has area three times that of the overlapping circular segment cut from the rolling disk. This surprising result is extended to epicycloids (and hypocycloids), obtained by rolling a disk of radius r externally (or internally) around a fixed circle of radius R. The factor 3 is replaced by $(3 + 2r/R)$ for the epicycloid, and by $(3 - 2r/R)$ for the hypocycloid.

This leads to several interesting consequences. For example, for any cycloid, epicycloid, or hypocycloid, the area of one full arch exceeds that of one full cap by twice the area of the rolling disk. Other applications yield (again without integration) compact geometrically revealing formulas for areas of cycloidal radial and ordinate sets.

The results are also extended to trochoids, in which the rolling disk rolls around a more general smooth base curve.

2.1 INTRODUCTION

For the average lay person the word roulette means a gambling game, or perhaps a small toothed wheel that makes equally spaced perforations like those on sheets of postage stamps. In geometry a *roulette* is the locus of a point attached to a plane curve that rolls along a fixed base curve without slipping. A surprising number of classical curves can be generated as roulettes – the cycloid, cardioid, tractrix, catenary, parabola, and ellipse, to name just a few. If the rolling curve is a circle,

the roulette is called a *trochoid* (from the Greek word $\tau\rho o\chi o\varsigma$ for wheel). Both the cycloid and cardioid are examples of trochoids.

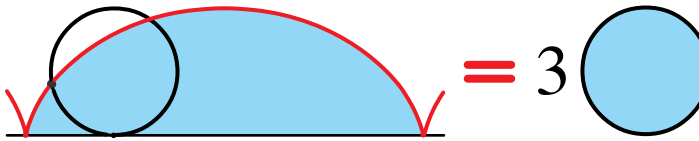


Figure 2.1: The cycloidal arch has area three times that of the rolling disk.

A *cycloid* is the path traced by a point on the boundary of a circular disk that rolls along a straight base line without slipping. The shaded region in Figure 2.1, obtained by one complete rotation of the disk, is called a *cycloidal arch*. It is known that its area is three times that of the rolling disk. Our analysis shows that the factor 3 reflects a much deeper property of cycloidal sectors.

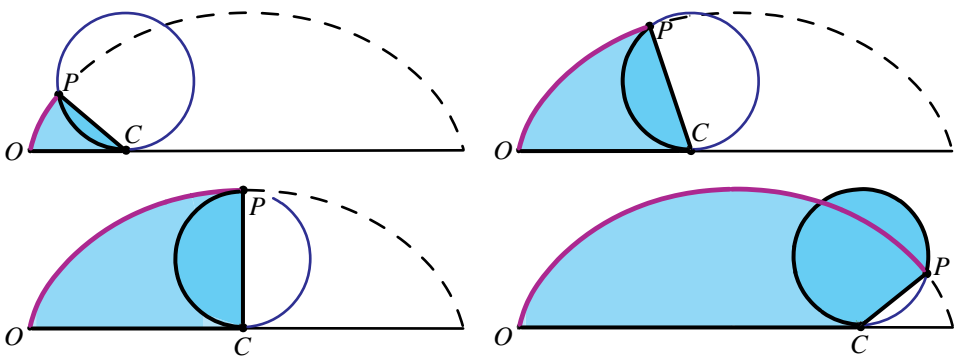


Figure 2.2: Each cycloidal sector OPC has area three times that of the shaded segment of the rolling disk cut by chord PC, where C is the contact point of the disk and base line.

Figure 2.2 displays various stages of the rotation of the disk, together with the *cycloidal sector* OPC bounded by arc OP and two line segments OC and PC. At each stage we have:

Theorem 2.1. *The cycloidal sector OCP has area three times that of the overlapping circular segment cut from the rolling disk by chord CP.*

This remarkable geometric property follows easily from an elegant area relation illustrated in Figure 2.3. The rolling disk is tangent to the upper and lower boundaries of the rectangle circumscribing the cycloid, at corresponding points of tangency T and C. The diameter TC divides the rolling circle into two semicircles, one of which intersects the cycloid at point P as indicated. In Figure 2.3, the line segment joining P and T cuts off a portion PCT of the rolling disk that we call a *wedge*. We call the region $PODT$, bounded by cycloidal arc PO, line segments OD, DT and TP, a *cycloidal cap*. The key that unlocks Theorem 2.1 is the following surprising relation:

Lemma 2.1. *The cycloidal cap PODT has the same area as the wedge PCT of the rolling disk.*

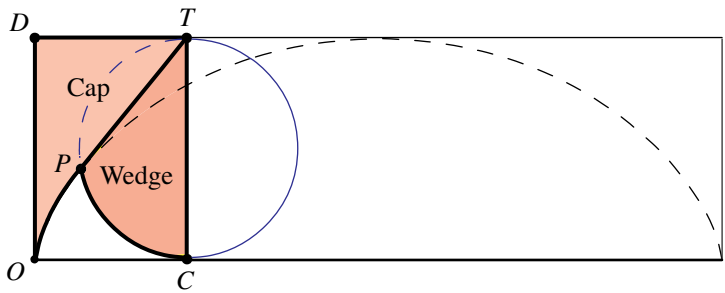


Figure 2.3: The cycloidal cap $PODT$ has the same area as the wedge PCT.

The foregoing area relations are extended to epicycloids and hypocycloids in Section 2.4, and to more general trochoids in Section 2.6.

2.2 AREA OF CYCLOIDAL CAP (PROOF OF LEMMA 2.1)

We deduce Lemma 2.1 as a consequence of Mamikon's sweeping-tangent theorem, which states that the area of a tangent sweep is equal to that of its corresponding tangent cluster. First we show that each chord PT is tangent to the cycloid at point P. This will show that as the disk rolls to the position shown in Figure 2.4, the tangent segment from the cycloid to the upper horizontal boundary of the rectangle, starting initially at OD, sweeps out the cycloidal cap $PODT$.

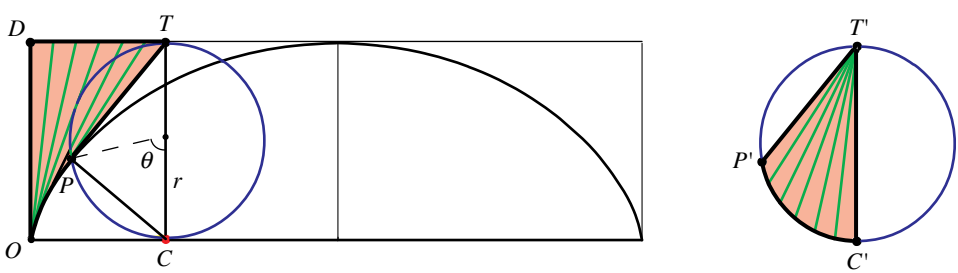


Figure 2.4: The cap, a tangent sweep, and its tangent cluster $T'C'P'$ have equal areas.

Instantaneous rotation principle.

To see that PT is tangent to the cycloid at P, note that triangle TPC is inscribed in the semicircle with diameter TC and hence is a right triangle. Because the disk rolls along the horizontal line without slipping, its point of contact C is instantaneously

at rest, and point P undergoes instantaneous rotation about C with PC as the instantaneous radius of rotation. This is called the *instantaneous rotation principle*, illustrated in Figure 2.5 for a disk bounded by a convex closed curve that rolls without slipping along a plane curve Γ.

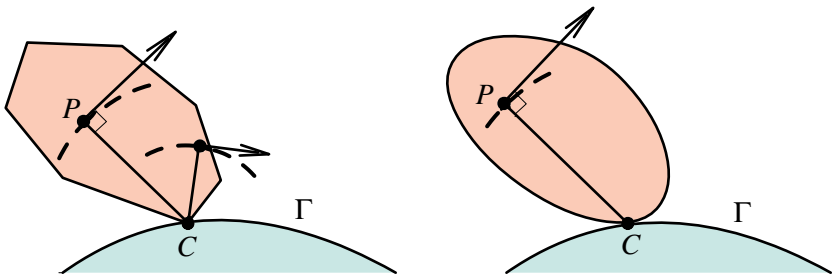


Figure 2.5: Instantaneous rotation principle. P undergoes instantaneous rotation about C, so PC is normal to the path of P.

The line segment joining an arbitrary point P on the disk to the point of contact C with Γ is *normal* to the path of P. Equivalently, a line through P perpendicular to PC (shown in Figure 2.5 with arrows) is *tangent* to the path of P. This principle is easily established for a polygon rolling about a vertex, and it holds more generally for all curves that are limits of polygons.

Apply the instantaneous rotation principle to the disk that generates the cycloid in Figure 2.4. Because the angle TPC is a right angle, the chord PT is perpendicular to the normal PC and hence is tangent to the cycloid. Thus, the cycloidal cap is a tangent sweep. To form the corresponding tangent cluster $T'C'P'$, translate each chord PT (parallel to itself) by moving all extremities T to one point T' on the right of Figure 2.4. Then the other extremity P moves to point P', with $P'T'$ equal in length and parallel to PT. Obviously, the segments $P'T'$ are chords of a circular disk congruent to the rolling disk. By Mamikon's theorem, the area of the tangent sweep $PODT$ is equal to that of the tangent cluster $T'C'P'$. The tangent cluster is congruent to the wedge TCP of the rolling disk in Figure 2.3, and we obtain Lemma 2.1.

2.3 AREA OF CYCLOIDAL SECTOR (PROOF OF THEOREM 2.1)

Throughout this chapter we employ square brackets to designate areas of regions. Thus, in Figure 2.6 we use the following notations:

[Sector] = area of the cycloidal sector OPC in Figures 2.2 and 2.6b.

[Tusk] = area of the tusk-like curvilinear region OPC below the cycloid and outside the disk (unshaded in Figure 2.6a).

[Wedge] = area of the wedge PCT of the circular disk (darker shading in Figure 2.6a).

[Segm] = area of the segment of the disk cut off by PC (darker shading in Figure 2.6b).

[Tri] = area of the right triangle TPC in Figure 2.6b.

[Rect] = area of the rectangle $ODTC$.

In these notations, Theorem 2.1 states that

$$[\text{Sector}] = 3[\text{Segm}]. \tag{2.1}$$

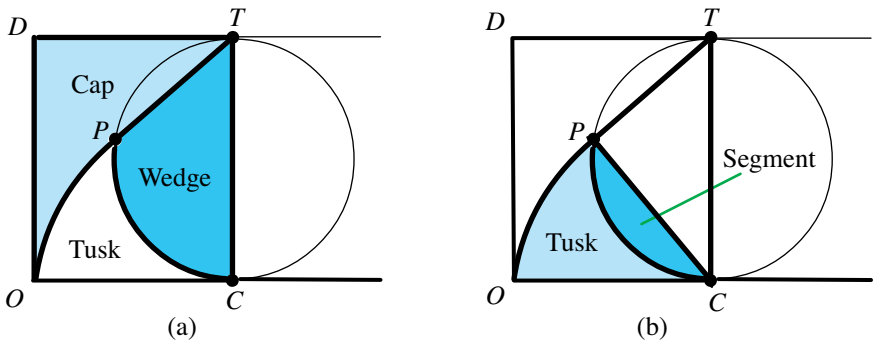


Figure 2.6: (a) The rectangle $ODTC$ is the union of Cap, Wedge, and Tusk. (b) The cycloidal sector OPC is the union of Tusk and Segment.

To begin the proof, Figure 2.6b reveals that [Sector] = [Segm] + [Tusk], so (2.1) is equivalent to

$$[\text{Tusk}] = 2[\text{Segm}].$$

Now by Lemma 2.1 the area of $PODT$ equals [Wedge], and from Figure 2.6a we find

$$[\text{Tusk}] = [\text{Rect}] - 2[\text{Wedge}] = [\text{Rect}] - 2[\text{Segm}] - 2[\text{Tri}], \tag{2.2}$$

where we used the relation

$$[\text{Wedge}] = [\text{Segm}] + [\text{Tri}]. \tag{2.3}$$

In Figure 2.4, OD has length $2r$, and OC has the length $r\theta$ of the circular arc CP, so

$$[\text{Rect}] = (2r)(r\theta) = 4(\frac{1}{2}r^2\theta). \tag{2.4}$$

From Figure 2.4 we see that $\frac{1}{2}r^2\theta$ is the area of the central sector of the disk subtended by central angle θ, which is also equal to [Segm] $+\frac{1}{2}$[Tri]. Therefore (2.4) implies

$$[\text{Rect}] = 4[\text{Segm}] + 2[\text{Tri}]. \tag{2.5}$$

When this is used in the right-hand side of (2.2), we obtain [Tusk] = 2[Segm], which proves Theorem 2.1.

Thus, the area of a cycloidal sector is three times that of a segment of the rolling disk. The area of the segment is given by the elementary formula

$$[\text{Segm}] = \frac{r^2}{2}(\theta - \sin\theta), \tag{2.6}$$

obtained by subtracting the area of the isosceles triangle with base PC in Figure 2.4 from the area of the circular sector with central angle θ.

Because the edge PC of a cycloidal sector POC is normal to the cycloid, as we observed in our discussion of instantaneous rotation, we see that POC is, in fact, swept by a moving segment normal to the cycloid.

2.4 EPICYCLOIDAL AND HYPOCYCLOIDAL CAP AND SECTOR

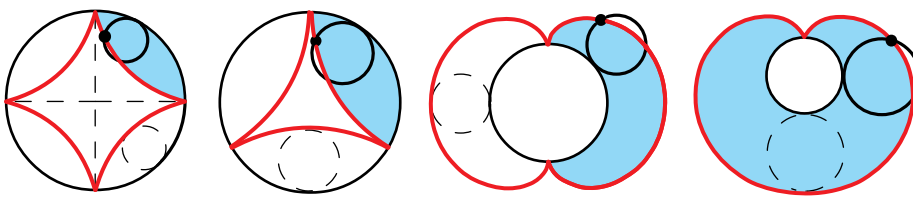


Figure 2.7: Examples of hypocycloids (a) and (b), and epicycloids (c) and (d).

Instead of rolling a circular disk along a fixed line we now roll it along the circumference of a fixed circle. Then a point on the boundary of the rolling disk traces a more general type of cycloid, called a *hypocycloid* if the disk rolls internally, with the two centers on the same side of the common tangent, as in Figures 2.7a and 2.7b, or an *epicycloid* if it is rolls externally, with the centers on opposite sides of the common tangent, as in Figures 2.7c and 2.7d. The shape will depend on the radius r of the rolling disk compared to the radius R of the fixed circle.

In Figure 2.7a, $r = R/4$ and the hypocycloid is called an *astroid*.

In Figure 2.7b, $r = R/3$ and the hypocycloid is called a *deltoid*.

The epicycloid in Figure 2.7c is a *nephroid* ($r = R/2$), and in Figure 2.7d it is a *cardioid* ($r = R$).

The cycloid in Figure 2.1 is the limiting case $R \to \infty$.

Extension of Theorem 2.1 to epicycloids and hypocycloids.

We extend Theorem 2.1 to both epicycloids and hypocycloids (Figure 2.8) by replacing the factor 3 by a new constant independent of the position of the rolling disk. In what follows, we assume that $r \leq R/2$ for hypocycloids, which entails no loss in generality because of a double generation theorem of Daniel Bernoulli that

ensures that the same family of epicycloids and hypocycloids is generated when $r > R/2$. The extended result is:

Theorem 2.2. *Every epicycloidal or hypocycloidal sector OPC has area $\omega_\pm$ times that of the overlapping segment of the rolling disk cut by chord PC, with $\omega_+ = 3 + 2r/R$ for the epicycloid, and $\omega_- = 3 - 2r/R$ for the hypocycloid.*

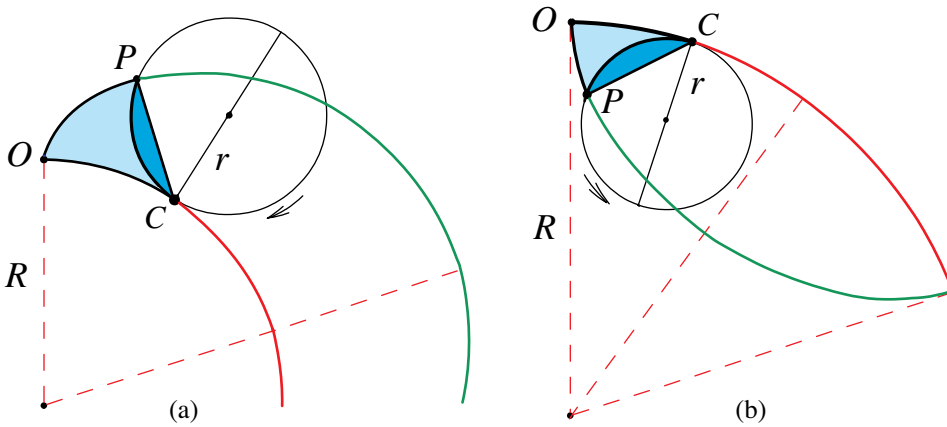


Figure 2.8: The sector OCP has area $\omega_\pm$ times that of the overlapping circular segment cut off by chord PC.

We will deduce Theorem 2.2 from Lemma 2.2, an extension of Lemma 2.1. In Lemma 2.1 the cycloidal cap $PODT$ and wedge PCT have equal areas, but in Lemma 2.2 (illustrated for an epicycloid in Figure 2.9a) the areas are related as follows:

Lemma 2.2. *Every epicycloidal or hypocycloidal cap $PODT$ has area $\kappa_\pm$ times that of wedge PCT of the rolling disk, with $\kappa_+ = 1 + 2r/R$ for the epicycloid, and $\kappa_- = 1 - 2r/R$ for the hypocycloid.*

Proof of Lemma 2.2. We treat the epicycloid first. Figure 2.9a shows an epicycloidal arc OP traced by a point P on a disk of radius r as it rolls along the outer circumference of a fixed circle of radius R. The epicyloid lies inside the annular ring between the fixed circle of radius R and the concentric circle of radius $R + 2r$. This ring plays a role similar to that of the rectangle circumscribing the cycloid in Figure 2.3.

Now refer to Figure 2.10a. The point P, initially at O, traces the portion OP of the epicycloid. The point of contact of the two circles, also initially at O, moves through an angle φ to point C as shown, tracing the circular arc OC of length $R\varphi$. The circular arc CP of radius r has length $r\theta$, where θ is the central angle in the rolling circle, as shown. Rolling takes place without slipping so the two circular arclengths CP and OC are equal:

$$r\theta = R\varphi. \tag{2.7}$$

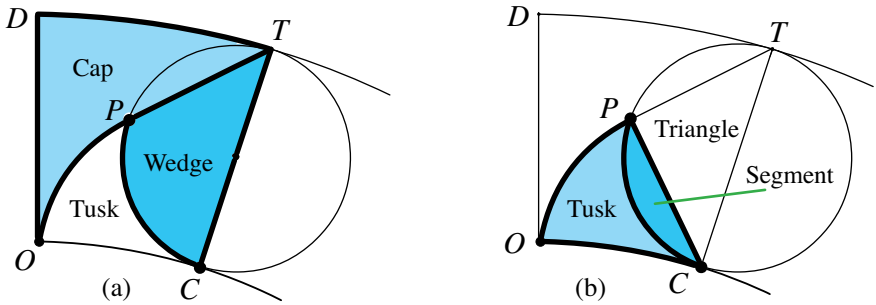


Figure 2.9: (a) Ring $ODTC$ is the union of epicycloidal Cap, Wedge, and Tusk. (b) Epicycloidal sector OPC is the union of Tusk and Segment.

A line through the two centers intersects the outer circle at its point of tangency T with the rolling circle. Triangle TPC is inscribed in a semicircle of diameter CT, so angle TPC is a right angle. The point C serves as the center of instantaneous rotation of the rolling circle, so PC is normal and PT is tangent to the epicycloid. As the tangent moves from its initial position OD to that shown in Figure 2.10a, it sweeps out the cycloidal cap $PODT$, turning through an angle

$$\alpha = \beta + \varphi. \tag{2.8}$$

Here β is the inscribed angle PTC subtending arc CP, and is half the central angle θ. From (2.7) we have $\varphi = r\theta/R = 2\beta r/R$, so (2.8) becomes

$$\alpha = \kappa_+ \beta, \tag{2.9}$$

where $\kappa_+ = 1 + 2r/R$. Form a tangent cluster by translating each tangent segment PT (parallel to itself) so point T moves to a fixed point T' in Figure 2.10b. By Mamikon's theorem, the area of the tangent sweep in Figure 2.10a is equal to that of the tangent cluster in Figure 2.10b. The tangent cluster is a portion of a rosette because the length of PT is equal to $2r\cos\beta = 2r\cos(\alpha/\kappa_+)$. The rosette area is equal to that of the circular wedge in Figure 2.10c multiplied by the factor κ_+, because the tangent cluster can be compressed to form the circular wedge in Figure 2.10c by rotating each translated segment $P'T'$ about T' to decrease the angle α by a factor of κ_+, from α to $\alpha/\kappa_+ = \beta$. But the wedge $T'C'P'$ in Figure 2.10c is congruent to the wedge TCP of the rolling disk in Figure 2.10a, so this proves Lemma 2.2 for the epicycloid.

A similar proof works for a hypocycloidal cap $PODT$ in Figure 2.11a. In this case $\alpha = \beta - \varphi$, giving us $\kappa_- = 1 - 2r/R$.

Proof of Theorem 2.2. We adapt the notation of Section 2.3 to Figure 2.9. Then Theorem 2.2 states that

$$[\text{Sector}] = \omega_+[\text{Segm}]. \tag{2.10}$$

But [Sector] = [Tusk] + [Segm], and $\omega_+ = \kappa_+ + 2$, so (2.10) is equivalent to

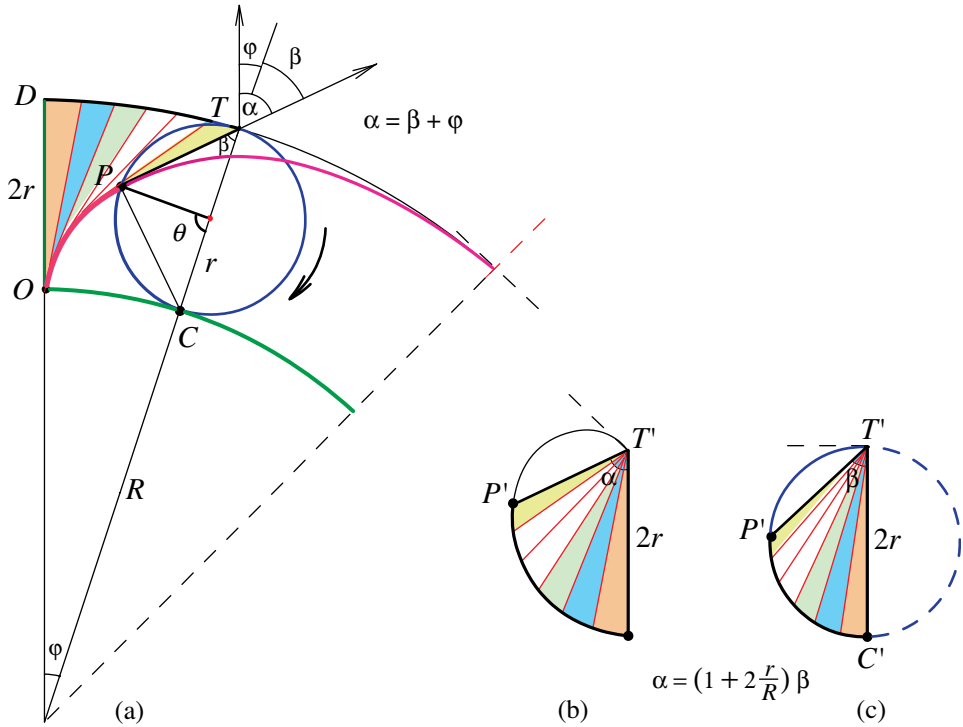


Figure 2.10: Proof of Lemma 2.2 for an epicycloidal sector.

$$[\text{Tusk}] = (\kappa_+ + 1)[\text{Segm}]. \tag{2.11}$$

To prove (2.11), refer to Figure 2.9a and use Lemma 2.2 and (2.3) to get

$$[\text{Tusk}] = [\text{Ring}] - [\text{Cap}] - [\text{Wedge}] = [\text{Ring}] - (\kappa_+ + 1)([\text{Segm}] + [\text{Tri}]). \tag{2.12}$$

Here [Ring] is the area of the portion $ODTC$ of the annular ring between the circles of radii R and $R + 2r$. On the other hand, $[\text{Ring}] = 2\varphi r(R + r) = 2r^2\theta(1 + r/R) = [\text{Rect}](\kappa_+ + 1)/2$ by (2.7) and (2.5), where $[\text{Rect}] = 2r^2\theta$ is the area of the rectangle $OCTD$ in Figure 2.3. From (2.5) we find

$$[\text{Ring}] = (\kappa_+ + 1)(2[\text{Segm}] + [\text{Tri}]).$$

Use this in the right-hand side of (2.12) to obtain (2.11). But (2.11) is equivalent to (2.10), hence this proves Theorem 2.2 for the epicycloid.

The same analysis based on Figure 2.11 works for the hypocycloid with the factors $\kappa_- = 1 - 2r/R$ and $\omega_- = 2 + \kappa_-$, because in this case

$$\alpha = \beta - \varphi.$$

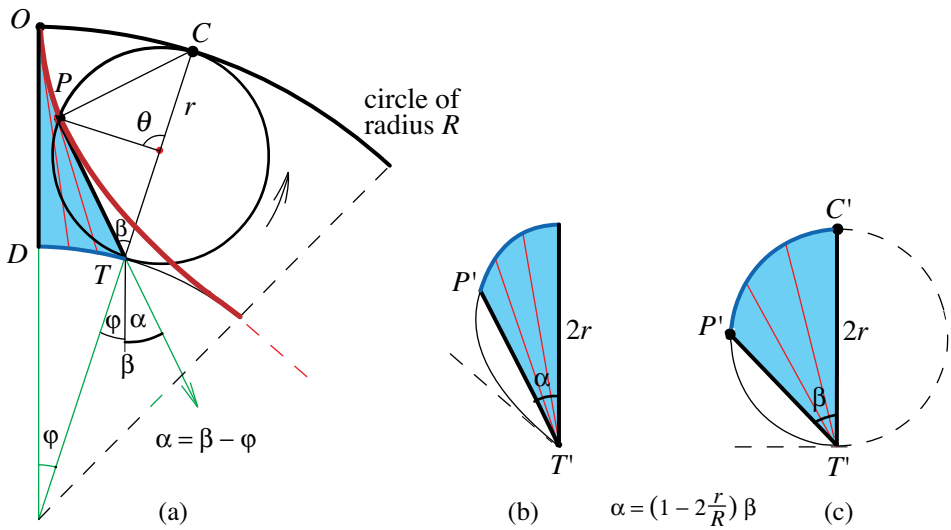


Figure 2.11: Proof of Lemma 2.2 for a hypocycloidal sector.

Area of full cap and full arch.

When a rolling disk D of area $[D]$ makes one complete rotation, the cycloidal, epicycloidal, or hypocycloidal sector fills a region we call a *full arch*. The corresponding cycloidal cap is called a *full cap*. From Lemma 2.2 and Theorem 2.2 we obtain:

Corollary 2.1. *For every epicycloid, the area of a full cap is $\kappa_+[D]$ and the area of a full arch is $\omega_+[D]$. For a hypocycloid, the respective areas are $\kappa_-[D]$ and $\omega_-[D]$, and for a cycloid they are $[D]$ and $3[D]$, where $\kappa_\pm = 1 \pm 2r/R$, and $\omega_\pm = \kappa_\pm + 2 = 3 \pm 2r/R$.*

The table gives values of r/R, $\kappa_\pm$, and $\omega_\pm$ for a few classical curves, including those in Figure 2.7. The first row relates to the cycloid, corresponding to $R = \infty$.

Curve	r/R	κ_+ or κ_-	ω_+ or ω_-
Cycloid	0	1	3
Cardioid	1	$\kappa_+ = 3$	$\omega_+ = 5$
Nephroid	1/2	$\kappa_+ = 2$	$\omega_+ = 4$
Deltoid	1/3	$\kappa_- = 1/3$	$\omega_- = 7/3$
Astroid	1/4	$\kappa_- = 1/2$	$\omega_- = 5/2$
Diameter	1/2	$\kappa_- = 0$	$\omega_- = 2$

The next two rows refer to two epicycloids with $\kappa_+ = 1 + 2r/R$: the cardioid ($r = R$) and the nephroid ($r = R/2$). The next two rows refer to two hypocycloids

with $\kappa_- = 1 - 2r/R$: the deltoid ($r = R/3$) and the astroid ($r = R/4$). The last entry with $r = R/2$ may seem unusual: this hypocycloid is the diameter of the fixed circle. The region between the arch and the circle of radius R is a semicircular disk of area $2\pi r^2$, half that of the circle.

Apparently no one has previously investigated the area of a cycloidal cap, but the result for a full arch is known. It is also derived in Chapter 3 by a different method (again without integration) in which both the rolling disk and the fixed circle are obtained as limits of regular polygons.

Because $\omega_\pm = \kappa_\pm + 2$, Corollary 2.1 also yields the following property (not previously recorded) that is common to all types of cycloids:

Corollary 2.2. *For any cycloid, epicycloid, or hypocycloid, the area of one full arch exceeds that of one full cap by twice the area of the rolling disk.*

Examples are shown in Figure 2.12, where the darker shaded region is the full arch and the lighter shaded region is a full cap. The difference in shaded areas is twice that of the rolling disk. If this general property could be established in a simpler way, then finding the areas of a full arch and a full cap would be almost trivial, because their sum is simply the area of a rectangle or annular ring.

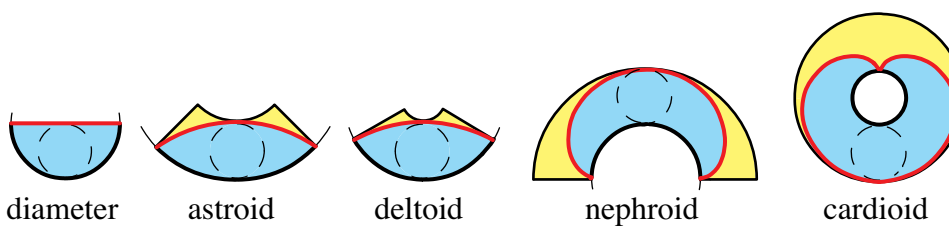


Figure 2.12: The difference in areas of shaded regions is twice the area of the rolling disks.

Another interesting property is inherited from the relation $\omega_+ + \omega_- = 6$. For every epicycloid obtained by a disk of radius r rolling outside a fixed circle of radius R, there is a corresponding hypocycloid obtained by a disk of the same radius r rolling inside, and we call the two cycloidal curves *complementary*. When the outside disk and the inside disk turn through the same angle θ, Theorem 2.2 tells us that the sum of the areas of the complementary sectors in Figure 2.8 is six times that of the overlapping circular segment cut from the rolling disk by chord PC, regardless of R. This implies the following property of the combined area of complementary full arches.

Corollary 2.3. *The sum of the areas of one full arch of an epicycloid and its complementary hypocycloid is equal to six times the area of the rolling disk.*

Examples are shown in Figure 2.13 with various values of R, but with a rolling disk of given radius r. In each case the sum of the darker and lighter shaded areas is six times that of the rolling disk. Incidentally, their difference is $(4r/R)[D]$.

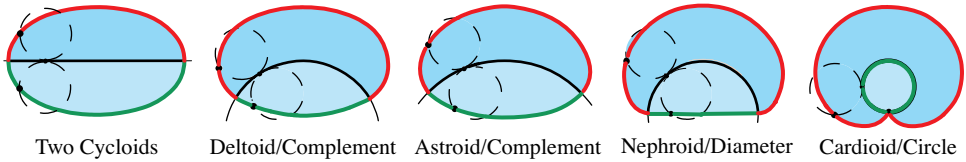


Figure 2.13: All combined regions have the same area: six times that of the rolling disk.

2.5 AREAS OF CYCLOIDAL RADIAL AND ORDINATE SETS

The area of an epicycloidal radial set (darker shading in Figure 2.14a) can be calculated by standard integration techniques based on a polar equation for the tracing point P involving the parameter θ of the rolling disk. The calculations are lengthy and the resulting formula is extremely complicated. This section shows how to find the area as a simple consequence of Theorem 2.2, avoiding polar equations and integration.

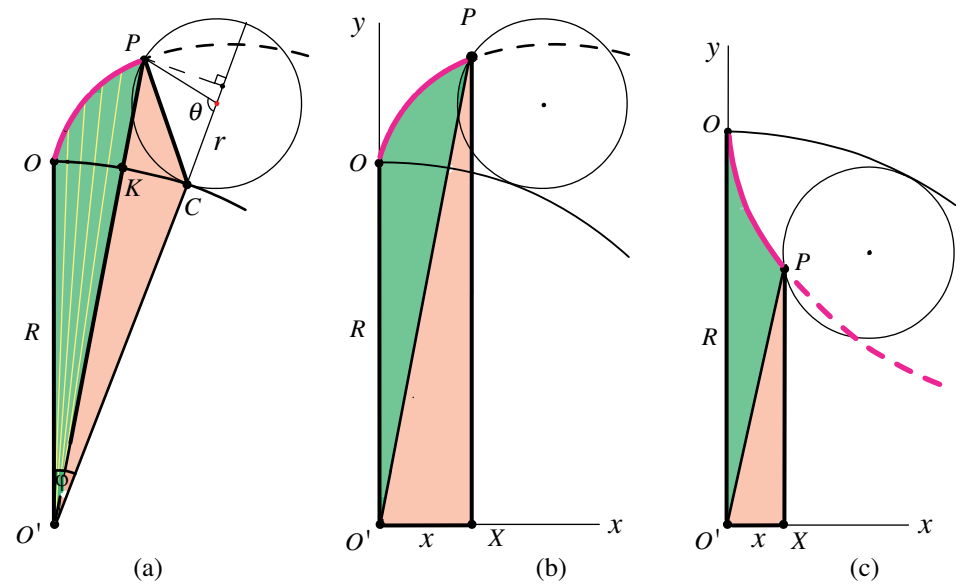


Figure 2.14: Geometric determination of the area of (a) an epicycloidal radial set $O'OP$, (b) an epicycloidal ordinate set $O'OPX$, and (c) a hypocycloidal radial and ordinate sets.

Consider the epicycloidal radial set $O'OP$ in Figure 2.14a. It is bounded by two radial segments $O'O$, $O'P$, and the epicycloidal arc OP. Denote its area by [Radial]. Then we have

$$[\text{Radial}] = [OPC] + [O'OC] - [O'PC], \tag{2.13}$$

where $[OPC]$ is the area of the epicycloidal sector OPC, $[O'OC]$ is the area of the

circular sector $O'OC$, and $[O'PC]$ the area of the triangle $O'PC$. Theorem 2.2 tells us that $[OPC] = \omega_+[\text{Segm}]$, where $\omega_+ = 3 + 2r/R$ and [Segm] is given by (2.6). Also, $[O'OC] = \frac{1}{2}R^2\varphi = \frac{1}{2}Rr\theta$, and $[O'PC] = \frac{1}{2}Rr\sin\theta$, so (2.13) can be written as

$$[\text{Radial}] = (\frac{R}{r} + \omega_+)[\text{Segm}]. \tag{2.14}$$

A corresponding formula holds for the area of a hypocycloidal radial set:

$$[\text{Radial}] = (\frac{R}{r} - \omega_-)[\text{Segm}].$$

Both results are contained in the following theorem.

Theorem 2.3. *The area of an epicycloidal or hypocycloidal radial set is given by*

$$[\text{Radial}] = (\frac{R}{r} \pm \omega_\pm)[\text{Segm}] = (\frac{R}{r} + 2\frac{r}{R} \pm 3)\frac{r^2}{2}(\theta - \sin\theta), \tag{2.15}$$

with the + sign for the epicycloid, and the − sign for the hypocycloid.

Now we can easily find the area [Ordinate] of the ordinate set $O'OPX$ in Figure 2.14b or 2.14c by adding the area of the right triangle $O'XP$ to that of the epicycloidal or hypocycloidal radial set $O'OP$:

Theorem 2.4. *The area of an epicycloidal or hypocycloidal ordinate set is given by*

$$[\text{Ordinate}] = (\frac{R}{r} \pm \omega_\pm)[\text{Segm}] + \frac{1}{2}xy \tag{2.16}$$

with the + sign for the epicycloid, and the − sign for the hypocycloid.

In (2.16), x and y are the rectangular coordinates of P with respect to the origin O' in Figure 2.14b or 2.14c. They are given in terms of θ by the well known parametric equations (which can be derived directly from Figure 2.14):

$$x = (R \pm r)\sin\frac{r\theta}{R} - r\sin(1 \pm \frac{r}{R})\theta, \quad y = (R \pm r)\cos\frac{r\theta}{R} \mp r\cos(1 \pm \frac{r}{R})\theta.$$

Here the upper sign is used for the epicycloid and the lower sign for the hypocycloid. The area of $O'OPX$ can also be calculated directly by integration, using the parametric equations. The calculation is tedious and leads to a lengthy unpleasant formula, in contrast to the compact and geometrically revealing (2.16).

As $R \to \infty$ in Figure 2.14b, the vertical segments OO' and PX remain parallel, and the portion of the region $O'OPX$ outside the circle of radius R becomes the shaded region shown in Figure 2.15a. This is the ordinate set of a cycloid, whose area we denote by $B(\theta)$. It is usually calculated by integration, using parametric equations for the cycloid.

We can also calculate $B(\theta)$ by a limiting argument based on (2.15) or (2.16), involving an indeterminate form of the type $\infty - \infty$. But we prefer to determine $B(\theta)$ directly and more simply from Theorem 2.1, avoiding parametric equations, integration, and indeterminate forms.

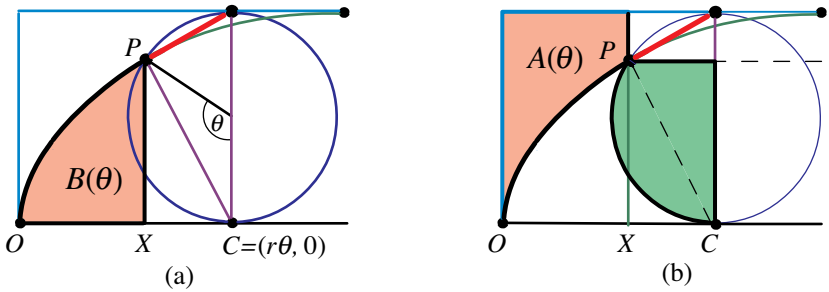


Figure 2.15: (a) The cycloidal sector OPC is the union of the ordinate set OXP and the triangle CXP. (b) The two shaded regions have equal areas.

To do so, note that in Figure 2.15a the area $B(\theta)$ equals the area of the cycloidal sector OPC minus the area $[CXP]$ of the right triangle CXP. Hence by Theorem 2.1 we have

$$B(\theta) = 3[\text{Segm}] - [CXP]. \tag{2.17}$$

In terms of the angle of rotation θ and the coordinates (x, y) of point P, we have $[CXP] = \frac{1}{2}y(r\theta - x)$, and [Segm] is given by (2.6).

Figure 2.15b reveals new area relations of interest. Let $A(\theta)$ denote the area of the shaded region directly above the ordinate set and inside the circumscribing rectangle. Using Lemma 2.1 it is easy to verify that $A(\theta)$ is equal to the area of the shaded portion of the rolling disk shown in Figure 2.15b, giving us

$$A(\theta) = [\text{Segm}] + [CXP]. \tag{2.18}$$

From (2.17) and (2.18) we see that

$$A(\theta) + B(\theta) = 4[\text{Segm}]. \tag{2.19}$$

In other words, the area of the rectangle with base OX and altitude $2r$ is always equal to 4[Segm]. This agrees with (2.5).

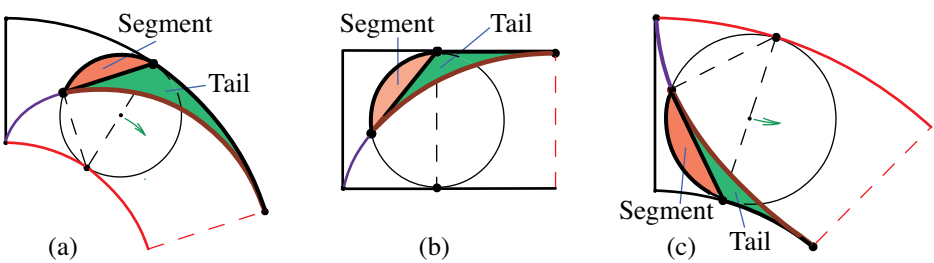


Figure 2.16: The area of (a) epicycloidal, (b) cycloidal, or (c) hypocycloidal tail is $\kappa_{\pm}$ times that of the adjacent segment of the rolling disk.

The equality of the areas of the shaded regions in Figure 2.15b also follows from the equality of the areas of the two shaded regions in Figure 2.16b, labeled "Segment" and "Tail." Again, Lemma 2.1 shows that for a cycloid we have [Tail] = [Segm]. For an epicycloidal or hypocycloidal tail, in Figures 2.16a and 2.16c, Lemma 2.2 leads to the result

$$[\text{Tail}] = \kappa_{\pm}[\text{Segm}]. \tag{2.20}$$

2.6 AREA OF A GENERAL TROCHOIDAL CAP AND SECTOR

Now we replace the fixed circle of radius R in Figure 2.8 by a more general piecewise smooth base curve Γ, along which we roll a disk of radius r. The curve Γ can have one or more points of inflection. In Figure 2.17a there is one such point B, where a change in the direction of bending occurs .

When the centers of curvature of the disk and Γ are initially on opposite sides of the common tangent at O, as in Figure 2.17a, we say that the disk rolls *externally* to Γ. A point P on the boundary of the disk traces a curve called a *epitrochoid*, generalizing the concept of epicyloid.

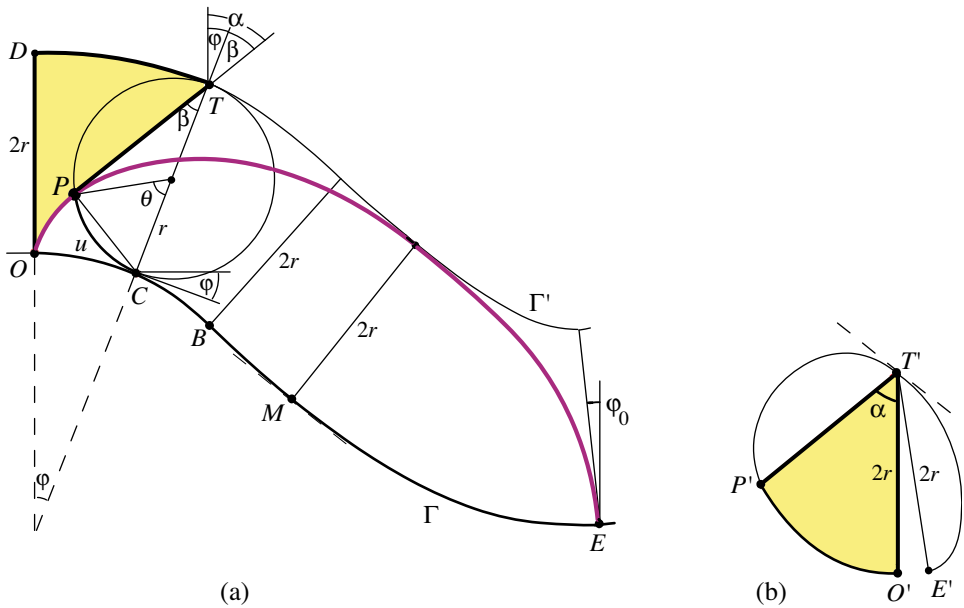


Figure 2.17: Diagram for determining the areas of an epitrochoidal cap and sector.

An epitrochoidal arch is generated after one revolution that begins with P at O and ends with P at E. The rolling disk in Figure 2.17a is also tangent to the curve Γ' parallel to Γ at a constant distance $2r$ from Γ. The parallel curves are the upper and lower boundaries of a curvilinear trapezoid that circumscribes the epitrochoidal

arch, analogous to the rectangle in Figure 2.3 circumscribing a cycloidal arch, or the portion of the circular ring in Figure 2.9a circumscribing the epicycloidal arch.

The rolling disk is initially tangent to Γ at O with OD as diameter. When the disk turns through an angle θ to the position shown in Figure 2.17a, the diameter CT through C makes an angle with the initial position OD denoted by φ. As before, TPC is a right triangle, PC is normal, and PT is tangent to the epitrochoid at P, $\beta = \theta/2$, and $\alpha = \beta + \varphi$ is the angle through which the epitrochoid tangent turns from its initial position OD. The angle φ increases as C moves along Γ from O to B, then decreases as C moves beyond B. After one revolution, P is at endpoint E and the diameter ET of the disk makes a final angle φ_0 with its initial position OD. When the disk rolls along Γ from O to B, the centers of curvature of the disk and of Γ are on opposite sides of the common tangent at C, and when it rolls from B to E the centers of curvature are on the same side of the common tangent. There is no restriction on the two radii of curvature.

If the centers of curvature of the disk and Γ are initially on the same side of the common tangent at O, we say the disk rolls *internally* to Γ, and point P traces a *hypotrochoid*, a generalization of a hypocycloid. The term *trochoid* refers to either an epitrochoid or a hypotrochoid. Our analysis deals primarily with areas relating to epitrochoids, and we indicate how the results need to be changed for hypotrochoids.

We shall extend both Theorems 2.1 and 2.2 to a general trochoid. Theorem 2.5a (stated below) relates the area of epitrochoidal cap $PODT$ in Figure 2.17a with the area [Wedge] of wedge TCP of the rolling disk. Theorem 2.5b relates the area of epitrochoidal sector POC with the area [Segm] of the segment of the rolling disk cut off by chord PC. We obtain these relations by comparing areas of the epitrochoidal cap and sector with those of the corresponding cycloidal cap and sector in Figure 2.3 obtained when a rolling disk of the same radius turns through the same angle θ. Corollary 2.4 relates the area of a full epitrochoidal cap and a full epitrochoidal arch with that of the rolling disk. Theorems 2.6 and Corollary 2.5 give corresponding results for hypothrochoids. As might be expected, the results for a general trochoid are more complicated than those in Theorem 2.1 (where Γ is a straight line), or in Theorem 2.2 (where Γ is a fixed circle).

To facilitate the comparisons in Theorems 2.5 we introduce the following notations that help relate areas of the general epitrochoid in Figure 2.17a to those of the cycloid in Figure 2.3. It is understood that a rolling disk of radius r rotates through the same angle θ in both cases, with the contact point moving along the arc OC, whose length is $r\theta$, the same as the length of circular arc PC.

[Trapez] = area of the curvilinear trapezoid $ODTC$ in Figure 2.17a.

[EpiSect] = area of the sector POC bounded by the line segment PC, the epitrochoidal arc OP, and the arc OC on Γ in Figure 2.17a.

[EpiCap] = area of the epitrochoidal cap $PODT$ above the epitrochoid in Figure 2.17a.

Our comparisons will be deduced from the following relation in Figure 2.17a:

$$[\text{EpiSect}] = [\text{Trapez}] - [\text{EpiCap}] - [\text{Tri}], \tag{2.21}$$

where, as before, [Tri] is the area of the right triangle TPC. We treat the first two terms on the right of (2.21) separately.

The area [Trapez] is equal to the average arclength of the two parallel boundaries times the distance $2r$ between them. Denote the length of the inner arc OC on Γ by u, which is a function of angle φ in Figure 2.17a, say $u = u(\varphi)$. Then the length of the outer arc DT on Γ' is $u + 2r\varphi$ so their average is $u + r\varphi$. When this is multiplied by the constant distance $2r$ between the curves we find that [Trapez] $= 2ru + 2r^2\varphi$. But $u = r\theta$ so $2ru = 2r^2\theta =$ [Rect], the area of $OCTD$ in Figure 2.6. Hence

$$[\text{Trapez}] = [\text{Rect}] + 2r^2\varphi. \tag{2.22}$$

Next we treat [EpiCap], the area of a tangent sweep. Its tangent cluster (with the same area) is bounded by two loops, shown in Figures 2.17b. It is obtained by translating each tangent segment PT (parallel to itself) so T moves to a fixed point T' and P moves to P'. The first loop is formed when the disk rolls along the first half of arc OE on Γ. The second loop is formed after the disk rolls past the midpoint M of arc OE. Let t denote the length of tangent segment PT in Figure 2.17a. The point P' has polar coordinates (t, α) taken with respect to T' as origin, where α is the angle of rotation of the tangent segment as indicated in Figure 2.17a. The area [EpiCap] is equal to the area of the corresponding tangent cluster, which is given by the following integral in polar coordinates:

$$[\text{EpiCap}] = \frac{1}{2}\int t^2\, d\alpha. \tag{2.23}$$

For simplicity in notation, we have written this as an indefinite integral, but it is actually a definite integral from 0 to α that would normally be expressed as $\int_0^\alpha t^2 d\alpha'$ with a dummy variable α' representing the angle the tangent makes with its initial position OD as the tracing point moves from O to P.

Recall that $\alpha = \beta + \varphi$, so $d\alpha = d\beta + d\varphi$, and (2.23) becomes

$$[\text{EpiCap}] = \frac{1}{2}\int t^2\, d\beta + \frac{1}{2}\int t^2\, d\varphi. \tag{2.24}$$

Each integral on the right of (2.24) has a natural geometric interpretation. The first term is the area of a cycloidal cap, the contribution to the tangent sweep that would occur when Γ is a straight line, of fixed direction. This is equal to [Wedge], the area of wedge TCP of the rolling disk. The second term occurs because Γ changes its direction through angle φ. Thus we can write (2.24) as

$$[\text{EpiCap}] = [\text{Wedge}] + \frac{1}{2}\int t^2\, d\varphi. \tag{2.25}$$

Using this and (2.22) in (2.21) we obtain

$$[\text{EpiSect}] = [\text{Rect}] - [\text{Wedge}] - [\text{Tri}] + 2r^2\varphi - \frac{1}{2}\int t^2\, d\varphi.$$

By (2.3) and (2.5), the first three terms on the right give 3[Segm], where [Segm] is the area of the circular segment cut from the rolling disk by chord PC. Therefore the foregoing equation can be written as

$$[\text{EpiSect}] = 3[\text{Segm}] + 2r^2\varphi - \frac{1}{2}\int t^2\, d\varphi. \tag{2.26}$$

The last two terms in (2.26) are equal to $\frac{1}{2}\int n^2 d\varphi$, where

$$n^2 = (2r)^2 - t^2. \tag{2.27}$$

The geometric meaning of n^2 is revealed by the right triangle TPC in Figure 2.17a, which has hypotenuse of length $2r$ and one leg of length t. By the Pythagorean theorem, n^2 is the square of the length of PC, which is normal to the epitrochoid. Therefore (2.26) takes the form

$$[\text{EpiSect}] = 3[\text{Segm}] + \frac{1}{2}\int n^2\ d\varphi. \tag{2.28}$$

Consequently, (2.25) and (2.26) give us:

Theorem 2.5. (a) *The area of an epitrochoidal cap and the area of the wedge TCP of the rolling disk are related by*

$$[\text{EpiCap}] = [\text{Wedge}] + \frac{1}{2}\int t^2\, d\varphi. \tag{2.29}$$

(b) *The area of an epitrochoidal sector and the area of the segment of the rolling disk cut off by chord PC are related by*

$$[\text{EpiSect}] = 3[\text{Segm}] + \frac{1}{2}\int n^2 d\varphi. \tag{2.30}$$

Note that an epitrochoidal cap is swept by tangent segments to the epitrochoid, and an epitrochoidal sector is swept by normal segments to the epitrochoid. Using tangents and normals to treat cycloidal and epitrochoidal areas is more natural than using rectangular coordinates.

Applying Theorem 2.5 to a full cap and to a full arch we find:

Corollary 2.4. *The area* [FullEpiCap] *of a full epitrochoidal cap is*

$$[\text{FullEpiCap}] = [D] + \frac{1}{2}\int_0^{\varphi_0} t^2\, d\varphi, \tag{2.31}$$

and the area [FullEpiArch] *of a full epitrochoidal arch is*

$$[\text{FullEpiArch}] = 3[D] + \frac{1}{2}\int_0^{\varphi_0} n^2 d\varphi. \tag{2.32}$$

Their sum is

$$[\text{FullEpiCap}] + [\text{FullEpiArch}] = 4[D] + 2r^2\varphi_0. \tag{2.33}$$

The sum of integrals in (2.31) and (2.32) simplifies because of (2.27).

The sum on the left of (2.33) is the area of the curvilinear trapezoid that circumscribes the full epitrochoidal arch in Figure 2.17.

Corresponding results for hypotrochoids.

The analysis for hypotrochoids is completely analogous to that for epitrochoids. The main difference is that when the disk rolls internally to Γ, as in Figure 2.11 for the hypocycloid, the relation $\alpha = \beta + \varphi$ is replaced by $\alpha = \beta - \varphi$. The argument used to prove Theorems 2.5 now gives us:

Theorem 2.6. (a) *The area of a hypotrochoidal cap and the area of the wedge TCP of the rolling disk are related by*

$$[\text{HypoCap}] = [\text{Wedge}] - \frac{1}{2}\int t^2\, d\varphi. \tag{2.34}$$

(b) *The area of a hypotrochoidal sector and the area of the segment of the rolling disk cut off by chord PC are related by*

$$[\text{HypoSect}] = 3[\text{Segm}] - \frac{1}{2}\int n^2 d\varphi. \tag{2.35}$$

The corresponding results for a full hypocycloidal cap and a full hypocycloidal arch are given by:

Corollary 2.5. *The area* [FullHypoCap] *of a full hypotrochoidal cap is given by*

$$[\text{FullHypoCap}] = [D] - \frac{1}{2}\int_0^{\varphi_0} t^2\, d\varphi, \tag{2.36}$$

and the area [FullHypoArch] *of a full hypotrochoidal arch is given by*

$$[\text{FullHypoArch}] = 3[D] - \frac{1}{2}\int_0^{\varphi_0} n^2 d\varphi. \tag{2.37}$$

Their sum is

$$[\text{FullHypoCap}] + [\text{FullHypoArch}] = 4[D] - 2r^2\varphi_0. \tag{2.38}$$

In (2.29) through (2.38), the area of an epitrochoidal or hypotrochoidal object is expressed (as predicted) in terms of the area of the corresponding cycloidal object, plus a correction term due to the curvature of Γ.

Area relations independent of Γ.

When a disk of given radius rolls on opposite sides of the curve Γ through the same angle, the corresponding portions of the epicytrochoidal and hypotrochoidal arches are called *complementary*. By adding the results in Theorems 2.5 and 2.6 we find that the correction terms due to the curvature of Γ cancel, and we obtain the following remarkable consequences:

Theorem 2.7. (a) *The sum of the areas of an epitrochoidal cap and its complementary hypotrochoidal cap does not depend on Γ, and is twice the area of the corresponding cycloidal cap:*

$$[\text{EpiCap}] + [\text{HypoCap}] = 2[\text{CycloCap}]. \tag{2.39}$$

(b) *The sum of the areas of an epitrochoidal sector and its complementary hypotrochoidal sector does not depend on* Γ*, and is twice the area of the corresponding cycloidal sector:*

$$[\text{EpiSector}] + [\text{HypoSector}] = 2[\text{CycloSector}]. \tag{2.40}$$

Proof. To obtain (2.39), add (2.29) to (2.34) and use Lemma 2.1 in the form [Wedge] = [CycloCap]. To obtain (2.40), add (2.30) to (2.35) and use Theorem 2.1 in the form 3[Segment]= [CycloSector].

In particular, for full arches we have

Corollary 2.6. *The sum of the areas of a full epitrochoidal cap and its complementary full hypotrochoidal cap is equal to twice the area of the rolling disk:*

$$[\text{FullEpiCap}] + [\text{FullHypoCap}] = 2[D], \tag{2.41}$$

Corollary 2.7. *The area of a full epitrochoidal arch and its complementary full hypotrochoidal arch is six times the area of the rolling disk:*

$$[\text{FullEpiArch}] + [\text{FullHypoArch}] = 6[D]. \tag{2.42}$$

Special case: Centrally symmetric base curve Γ.

An example occurs in Figure 2.18, where the base curve Γ consists of two adjacent semicircles with their point of contact being a point of central symmetry for Γ. In

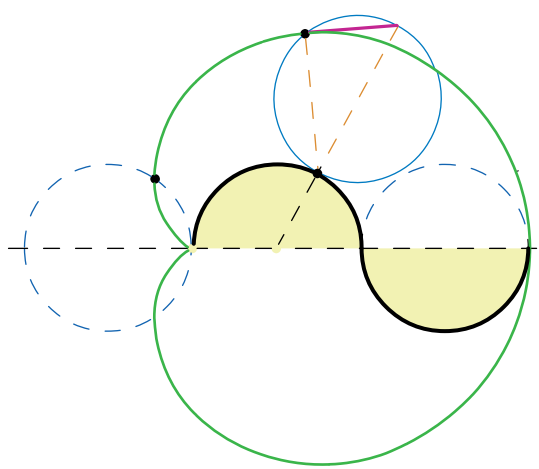


Figure 2.18: Epitrochoid used to form half a cardioid.

this example, the epitrochoid is half a cardioid. Here the disk D of radius r and area $[D] = \pi r^2$ rotates along a curve Γ composed of two adjacent semicircular arcs of radius r. One arch of the trochoid so generated consists of the region above Γ and below the cardioid. The area of the trochoidal arch is $3[D]$, because it consists

of the upper half of the cardioidal region, which we know (from the table) to be $\frac{5}{2}[D]$, plus the area of a semicircular disk of area $\frac{1}{2}[D]$.

The following general result for a centrally symmetric base curve is a direct consequence of Corollaries 2.6 and 2.7.

Corollary 2.8. *For a centrally symmetric base curve Γ, the area of a full epitrochoidal cap is $[D]$, and the area of the full epitrochoidal arch is $3[D]$, the same as for a cycloid.*

Corollary 2.8 is obtained by applying (2.41) and (2.42) to the regions shown in Figure 2.19. The epitrochoidal cap consists of two parts, C_1 and C_2, and the epitrochoidal arch consists of two parts A_1 and A_2, as indicated in Figure 2.19. Because of central symmetry, each of these four parts is congruent to a corresponding symmetric hypotrochoidal part below Γ having the same label. By (2.41) we find $[C_1]+[C_2] = [D]$, and by (2.42) we obtain $[A_1]+[A_2] = 3[D]$, which yields Corollary 2.8.

The difference of the two areas in Corollary 2.8 is $2[D]$, and their sum, which represents the area of the circumscribing curvilinear trapezoid, is $4[D]$, just as in the case of cycloidal, epicycloidal, and hypocycloidal arches.

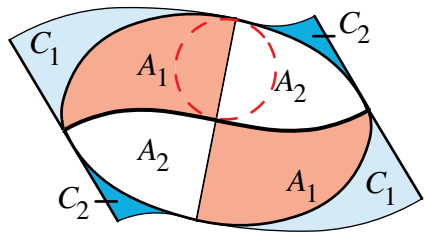


Figure 2.19: Proof of Corollary 2.8 for a centrally symmetric base curve.

Area formulas in terms of intrinsic description of Γ.

The area formulas for a full trochoidal arch can also be expressed in an alternative form. Recall that we have $t^2 = (2r)^2\cos^2\beta = 2r^2(1+\cos\theta)$ and $n^2 = (2r)^2\sin^2\beta = 2r^2(1-\cos\theta)$, hence the integrals in (2.31) and (2.32) become

$$\frac{1}{2}\int_0^{\varphi_0} t^2 d\varphi = r^2\varphi_0 + r^2 I(\varphi_0), \text{ and } \frac{1}{2}\int_0^{\varphi_0} n^2 d\varphi = r^2\varphi_0 - r^2 I(\varphi_0), \tag{2.43}$$

where

$$I(\varphi_0) = \int_0^{\varphi_0} \cos\theta\, d\varphi. \tag{2.44}$$

In (2.44) we express both θ and φ as functions of the arclength u of OC on the curve Γ shown in Figure 2.17a. The length of the arc OC on Γ is equal to $r\theta$, the length of the circular arc PC on the rolling circle. Hence $u = r\theta$, or $\theta = u/r$. We cannot express φ explicitly in terms of u until Γ is known, but we can indicate the

relation as $\varphi = \varphi(u)$, which gives an intrinsic description of Γ. Note that $\varphi(u) = \varphi_0$ when $u = 2\pi r$.

Integration by parts in the integral in (2.44) gives

$$\int_0^{\varphi_0} \cos\theta \, d\varphi = \varphi_0 + \int_0^{2\pi r} \varphi(u) \sin\frac{u}{r} d\frac{u}{r}.$$

When this is used in (2.43) the term $r^2\varphi_0$ cancels and we get

$$\frac{1}{2}\int_0^{\varphi_0} n^2 d\varphi = -r\int_0^{2\pi r} \varphi(u)\sin\frac{u}{r}du. \tag{2.45}$$

For the first integral in (2.43) the term $r^2\varphi_0$ does not cancel and we find

$$\frac{1}{2}\int_0^{\varphi_0} t^2 d\varphi = 2r^2\varphi_0 + r\int_0^{2\pi r} \varphi(u)\sin\frac{u}{r}du. \tag{2.46}$$

The formulas in Corollaries 2.4 and 2.5 now give us:

Theorem 2.8. *If $\varphi(u)$ is the intrinsic description of Γ, we have*

$$[\text{FullEpiCap}] = [D] + 2r^2\varphi_0 + r\int_0^{2\pi r} \varphi(u)\sin\frac{u}{r}du, \tag{2.47}$$

$$[\text{FullEpiArch}] = 3[D] - r\int_0^{2\pi r} \varphi(u)\sin\frac{u}{r}du, \tag{2.48}$$

$$[\text{FullHypoCap}] = [D] - 2r^2\varphi_0 - r\int_0^{2\pi r} \varphi(u)\sin\frac{u}{r}du, \tag{2.49}$$

$$[\text{FullHypoArch}] = 3[D] + r\int_0^{2\pi r} \varphi(u)\sin\frac{u}{r}du. \tag{2.50}$$

Special case: Circular base curve Γ.

For this special case, Theorem 2.2 follows from Theorem 2.5b. For an epicycloid, Γ is a circle of radius R, and the arclength $u = R\varphi = r\theta$, so $\varphi = r\theta/R$ and $d\varphi = (r/R)d\theta$. To deduce Theorem 2.2 we calculate the integral $\int n^2 d\varphi$ in (2.30) by noting that

$$n^2 = (2r)^2\sin^2\beta = 2r^2(1-\cos 2\beta) = 2r^2(1-\cos\theta).$$

Hence

$$\frac{1}{2}\int n^2 d\varphi = r^2\int(1-\cos\theta)\frac{r}{R}d\theta = \frac{2r}{R}\frac{r^2}{2}(\theta - \sin\theta) = \frac{2r}{R}[\text{Segm}],$$

where we have used (2.6) for [Segm] in Figure 2.6b. Thus (2.30) gives us

$$[\text{TroSect}] = (3 + \frac{2r}{R})[\text{Segm}] = \omega_+[\text{Segm}],$$

in agreement with Theorem 2.2. The result for a hypocycloid is obtained by noting that $d\varphi/d\theta$ is negative, so $d\varphi = -(r/R)\,d\theta$. Similarly, Lemma 2.2 follows from Theorem 2.5a.

2.7 NEW APPLICATIONS OF THEOREM 2.8

This section applies Theorem 2.8 to some famous curves that have not been previously used as base curves for generating trochoids.

Cornu spiral as base curve Γ.

This curve, also called a *clothoid*, was treated by Euler in 1781 in a study of an elastic spring. It is also involved in problems concerning the diffraction of light.

For our purposes, we define the base curve Γ by the intrinsic equation $\varphi(u) = cu^2$, where c is a positive constant. If a disk of radius r rolls along Γ, starting at the point where $\varphi = u = 0$ and making one complete turn, the integral in (2.48) with $\varphi(u) = cu^2$ is equal to

$$c\int_0^{2\pi r} u^2 \sin\frac{u}{r}du = cr^3\int_0^{2\pi} \theta^2 \sin\theta\, d\theta = -4\pi^2 cr^3 = -4\pi rc[D], \tag{2.51}$$

where $[D] = \pi r^2$ is the area of the rolling disk. Using this in (2.48) we find

$$[\text{FullEpiArch}] = 3[D] + 4c[D]^2.$$

The second term can also be written in terms of the parameter a that locates the poles of the spiral at $(\pm a, \mp a)$. It is known that $a^2c = \pi/8$, so the formula for the area of a full epitrochoidal arch now becomes

$$[\text{FullEpiArch}] = (3 + \frac{1}{2}(\frac{\pi r}{a})^2)[D].$$

The corresponding formula for a full hypotrochoidal arch is

$$[\text{FullHypoArch}] = (3 - \frac{1}{2}(\frac{\pi r}{a})^2)[D].$$

The example shown in Figure 2.20a has $r = a/\pi$, which gives $3\frac{1}{2}[D]$ for the area of the rightmost arch, and $2\frac{1}{2}[D]$ for the leftmost arch.

We were pleased to find that, for this value of r, the arclength of the spiral from the origin to the first vertical tangent is exactly $2a$. The difference of these two areas is $[D]$. If the same disk makes another complete clockwise turn beyond the vertical tangent the new trochoidal arch will have area $4\frac{1}{2}[D]$. Each further turn adds another $[D]$ to the area of the trochoidal arch.

When $r = a/2$ the factor multiplying $[D]$ is $(3 \pm \pi^2/8)$ giving us

$$[\text{FullEpiArch}] \approx 4.27[D], \text{ and } [\text{FullHypoArch}] \approx 1.73[D].$$

When $r = a\sqrt{6}/\pi$ we find

$$[\text{FullEpiArch}] = 6[D], \text{ and } [\text{FullHypoArch}] = 0.$$

Figure 2.20b shows why [FullHypoArch] = 0 in this case. The trochoid has a cusp at the point where the radius of the rolling disk is equal to the radius of curvature

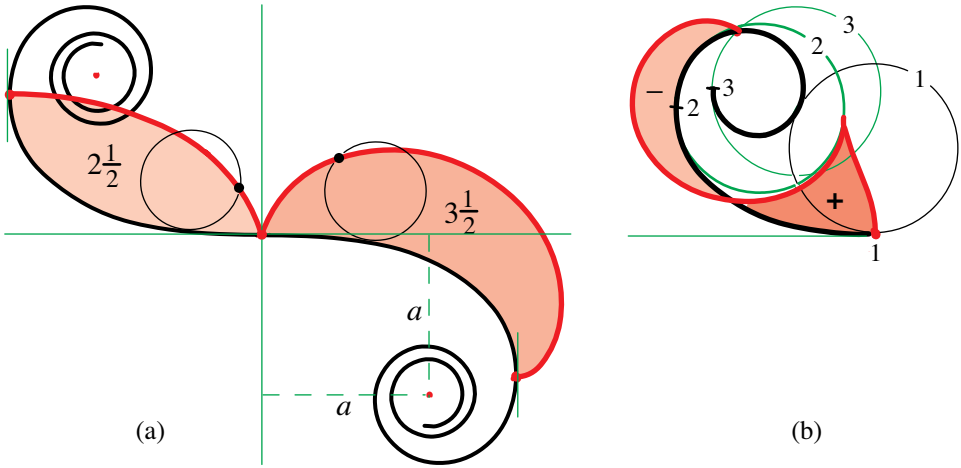


Figure 2.20: (a) Cornu spiral as base curve, with $\varphi(u) = cu^2$. The spiral winds around two poles with coordinates $(\pm a, \mp a)$, where $a^2c = \pi/8$. When $r = a/\pi$, the area of the epitrochoidal arch is $3\frac{1}{2}[D]$, and that of the hypotrochoidal arch is $2\frac{1}{2}[D]$. In (b), $r = a\sqrt{6}/\pi$, and the left portion of the trochoid forms a cusp and intersects the base curve as shown, forming two arches on opposite sides of Γ (denoted as $+$ and $-$) whose areas are equal.

of the base curve. At the cusp, the disk changes its direction of rolling, from counterclockwise to clockwise. Then the trochoid intersects the base curve forming two arches, one on the concave side of Γ, the other on the convex side, of equal areas. In general, [FullEpiArch] is positive, but [FullHypoArch] can be positive or negative. Their sum is always $6[D]$.

Logarithmic spiral as base curve Γ.

In this example, the arclength of Γ is given by $u = L(e^{a\varphi} - 1)$, (see Example 3 in Chapter 11) hence $\varphi = (1/a)\log(1 + u/L)$, and the integral in (2.48) becomes

$$\frac{1}{a}\int_0^{2\pi} \log(1 + \frac{r\theta}{L})\sin\theta\, d\theta. \tag{2.52}$$

The integral multiplying $1/a$ can be expressed in terms of the transcendental functions $\mathrm{Ci}(x)$ and $\mathrm{Si}(x)$, defined by the integrals

$$\mathrm{Ci}(x) = \gamma + \log x + \int_0^x \frac{\cos\theta - 1}{\theta} d\theta \quad \text{and} \quad \mathrm{Si}(x) = \int_0^x \frac{\sin\theta}{\theta} d\theta,$$

where γ is Euler's constant. A closed-form formula for the integral in (2.52) is $1/a$ times the quantity

$$\cos\frac{L}{r}(\mathrm{Ci}(\frac{L}{r} + 2\pi) - \mathrm{Ci}(\frac{L}{r})) + \sin\frac{L}{r}(\mathrm{Si}(\frac{L}{r} + 2\pi) - \mathrm{Si}(\frac{L}{r})) - \log(1 + \frac{2\pi r}{L}).$$

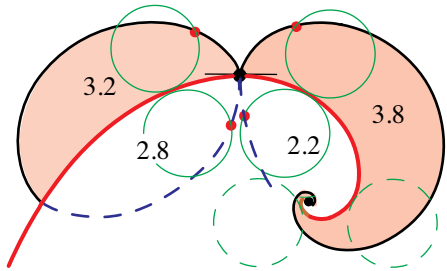


Figure 2.21: Logarithmic spiral as base curve, with $\varphi(u) = (1/a)\log(1 + u/L)$.

When $L = 2\pi r$ the foregoing quantity simplifies to

$$\mathrm{Ci}(4\pi) - \mathrm{Ci}(2\pi) - \log 2 = \int_{2\pi}^{4\pi} \frac{\cos\theta - 1}{\theta}\, d\theta.$$

This example is illustrated in Figure 2.21, for which [FullEpiArch]= $3.8[D]$ and [FullHypoArch] = $2.2[D]$ for the areas of the two righmost arches in Figure 2.21. A similar calculation gives the results [FullEpiArch]= $3.2[D]$ and [FullHypoArch] = $2.8[D]$ for the two leftmost arches in Figure 2.21.

Tractrix and catenary as base curves Γ.

For the tractrix as base curve, shown in Figure 2.22a, we have $u = -k\log(\cos\varphi)$ (see (11.12)), hence $\varphi(u) = \exp(-\arccos(u/k))$, and the integral in (2.48) takes the

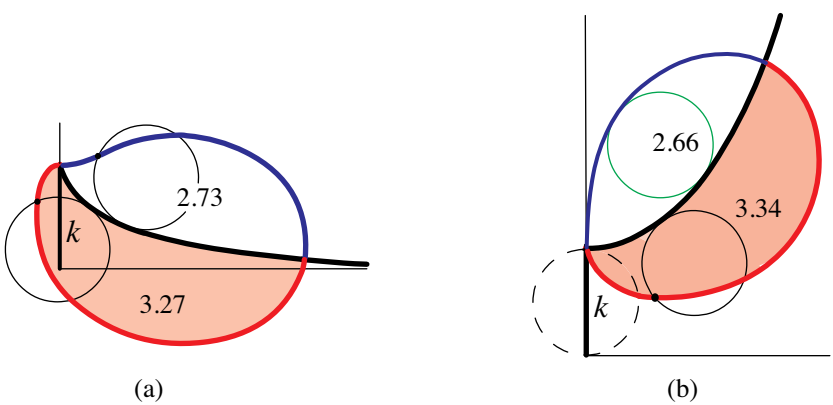


Figure 2.22: (a) Tractrix as base curve, with $\varphi(u) = \exp(-\arccos(u/k))$. (b) Catenary as base curve, with $\varphi(u) = \arctan(u/k)$.

form

$$\int_0^{2\pi r} \exp(-\arccos\frac{u}{k})\sin\frac{u}{r}du = r\int_0^{2\pi} \exp(-\arccos\frac{r\theta}{k})\,\sin\theta\, d\theta.$$

When $r = k$, the area of one epitrochoidal arch is $3.27[D]$, and that of the hypotrochoidal arch is $2.73[D]$, where $[D]$ is the area of the rolling disk.

For the catenary as base curve, shown in Figure 2.22b, we have $u = k \tan\varphi$ and the integral in (2.44) takes the form

$$\int_0^{2\pi r} \arctan\frac{u}{k} \sin\frac{u}{r} du = r \int_0^{2\pi} \arctan\frac{r\theta}{k} \sin\theta \, d\theta.$$

When $r = k$, the area of one epitrochoidal arch is $3.34[D]$, and that of the hypotrochoidal arch is $2.66[D]$.

Cycloid as base curve Γ.

This base curve is a cycloid, generated by a disk of radius R rolling along a horizontal line.

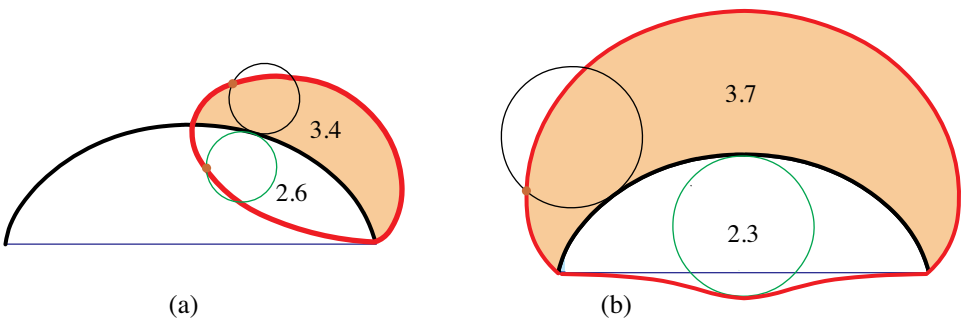


Figure 2.23: Cycloid generated by disk of radius R as base curve. In (a), $u = 4R\sin\varphi$, and in (b) $u = 4R(1 - \cos\varphi)$.

In Figure 2.23a, a disk of radius r rolls along the cycloid from its highest point, so that $u = 4R\sin\varphi$. The radii r and R are related by $2\pi r = 4R$, half the length of the cycloid. In Figure 2.23b, $u = 4R(1 - \cos\varphi)$, and a disk of radius r with $2\pi r = 8R$ rolls along the cycloid from one end to the other. In (a), the integral in (2.48) takes the form

$$\int_0^{2\pi r} \arcsin\frac{u}{4R} \sin\frac{u}{r} du = r \int_0^{2\pi} \arcsin\frac{\theta}{2\pi} \sin\theta \, d\theta.$$

and in (b) it becomes

$$\int_0^{2\pi r} \arccos(1 - \frac{u}{4R}) \sin\frac{u}{r} du = r \int_0^{2\pi} \arccos(1 - \frac{\theta}{\pi}) \sin\theta \, d\theta.$$

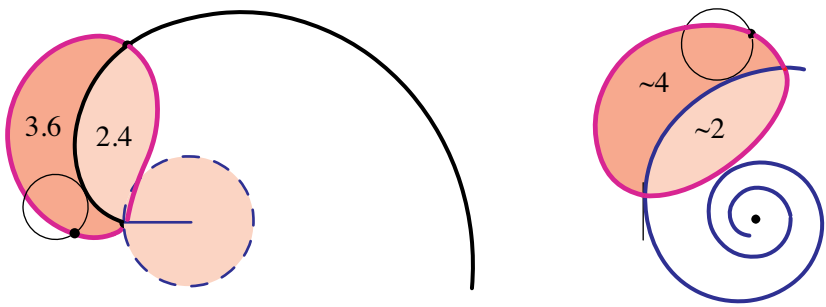
This yields the values $3.4[D]$ and $3.7[D]$ for the two epitrochoidal arches, and the values $2.6[D]$ and $2.3[D]$ for the hypotrochoidal arches.

Involute of a circle as base curve Γ.

Figure 2.24a shows the epitrochoid and hypotrochoid formed by rolling a circle of radius $r = a/2$ along the involute of a circle of radius a. In this case we have $u = r\varphi^2$ (see Section 11.11), hence $\varphi(u) = \sqrt{u/r}$ and the integral in (2.48) takes the form

$$\int_0^{2\pi r} \sqrt{\frac{u}{r}} \sin \frac{u}{r} du = r \int_0^{2\pi} \sqrt{\theta} \sin \theta \, d\theta.$$

This yields $3.6[D]$ and $2.4[D]$ for the areas of the corresponding arches.



(a) Involute of circle (b) Hyperbolic Spiral

Figure 2.24: (a) The base curve is the involute of a circle of radius a, and the rolling disk has radius $r = a/2$. (b) The base curve is a hyperbolic spiral with intrinsic description $\varphi(u) = 2\pi u/(a-u)$, and the rolling disk has radius $r = a/(6\pi)$.

Hyperbolic spiral as base curve Γ.

Figure 2.24b shows the epitrochoid and hypotrochoid formed by rolling a disk of radius $r = a/(6\pi)$ along a base curve Γ with arclength function $u(\varphi) = a\varphi/(\varphi+2\pi)$. Then $u(0) = 0$ and $u(\varphi) \to a$ as $\varphi \to \infty$. We call Γ a hyperbolic spiral; it has intrinsic description $\varphi(u) = 2\pi u/(a-u)$. For this choice of $\varphi(u)$ the value of the integral in (2.48) is very nearly equal to $-\pi r^2$, and this yields values nearly equal to $4[D]$ and $2[D]$ for the areas of the corresponding arches.

A challenging extremum problem.

The foregoing examples suggest an interesting question:

How large can the area of an epitrochoidal arch be when a convex base curve has a perimeter equal to that of the rolling disk?

Figure 2.25 shows the prototype of this problem. Here the base curve is a line segment, the trochoid is a cycloid, and the area of the cycloidal arch is exactly $3[D]$, where $[D]$ is the area of the rolling disk. To interpret the problem geometrically, imagine the cycloidal arch dissected into a large number of thin vertical slices, like the teeth of a comb, as suggested by Figure 2.25. The base of the comb has length

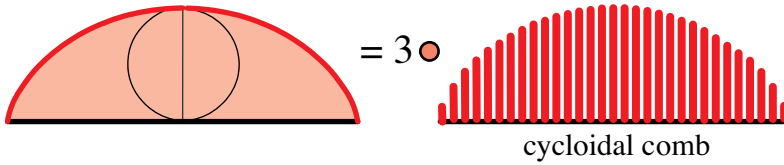


Figure 2.25: Dissecting a cycloidal arch.

equal to the perimeter of the rolling disk. The next three figures show the base of the comb wrapped along various base curves of the same perimeter. Figure 2.26 shows two such base curves, one obtained by folding the original base line in half

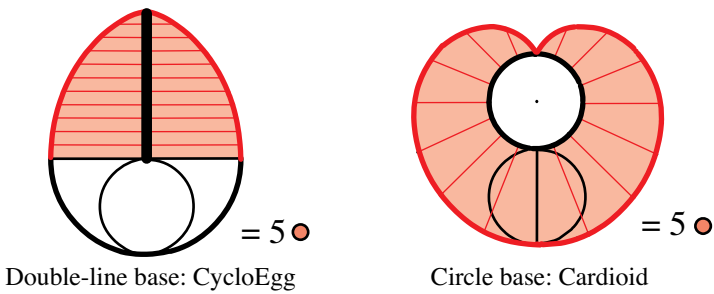


Figure 2.26: Baseline folded two different ways to form epitrochoidal arch of area $5[D]$.

to form a double base line as Γ, the second by wrapping the base line along a circle with the radius of the rolling disk to form a circular base curve Γ. In the two examples, the corresponding epitrochoidal arch has area exactly $5[D]$.

Figure 2.27 shows the original base line folded to form an equilateral triangle and a square, each with perimeter equal to the circumference of the rolling disk. The corresponding epitrochoidal arches have approximate areas $5.3[D]$ and $5.4[D]$, respectively. In Figure 2.28a, the original horizontal base line is divided into four

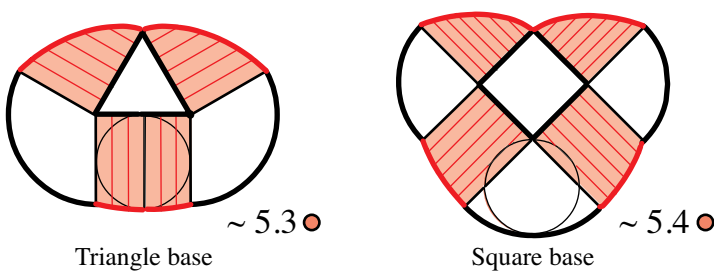
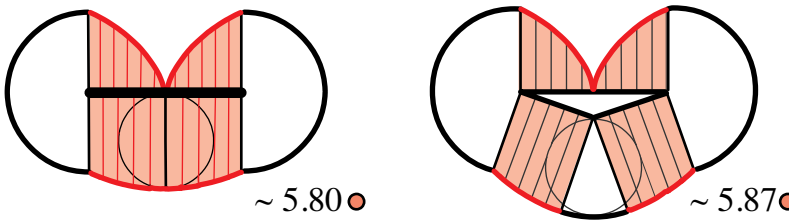


Figure 2.27: Baseline folded to form an equilateral triangle and a square.

equal parts and refolded to form another double base line. This time the disk is

rolled as indicated in the drawing to form an arch of area $5.80[D]$. A slightly larger area, $5.87[D]$ is obtained by using the triangular base shown in Figure 2.28b. These examples suggest that the maximum possible area of the epitrochoidal arch is $6[D]$. The problem can also be formulated analytically: We seek a convex base curve described intrinsically by $\varphi(u)$ that maximizes the integral $\int_0^{2\pi} \varphi(r\theta) \sin\theta \, d\theta$.



(a) Double-line base (b) Obtuse triangle base

Figure 2.28: Baseline folded to form a double line and an obtuse isosceles triangle.

2.8 SPECIAL RESULTS ON CYCLOIDAL AREA

Roberval and Toricelli were the first to calculate the area of a cycloidal arch. Roberval introduced a "companion" curve that, in fact, is a sine curve. The diagrams in Figure 2.29 use essentially the same idea but without the auxiliary sine curve.

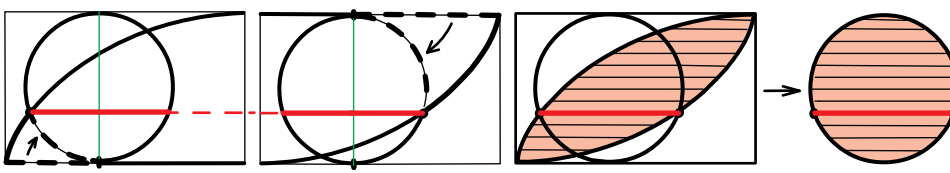


Figure 2.29: Half a cycloidal arch inscribed in a rectangle.

We leave it to the reader to explain how the diagrams provide a proof without words that the area of the shaded region is equal to that of the rolling disk. From this fact it is easy to deduce that the area of the region under a full cycloidal arch is three times that of the rolling disk.

Further charming contributions to the history of cycloidal quadrature by Huygens, Leibniz, and Johannis Bernoulli, are illustrated in Figure 2.30, where a rectangle circumscribes a cycloidal arch, the generating circular disk is at its center, and a horizontal line bisects the rectangle.

In Figure 2.30a, the upper half of the rectangle is bisected by a horizontal dashed line. In 1658, Huygens proved that the area of the shaded cycloidal segment cut off by the upper dashed line is equal to the area of half the inscribed hexagon, which we display as the area of the adjacent lower equilateral triangle.

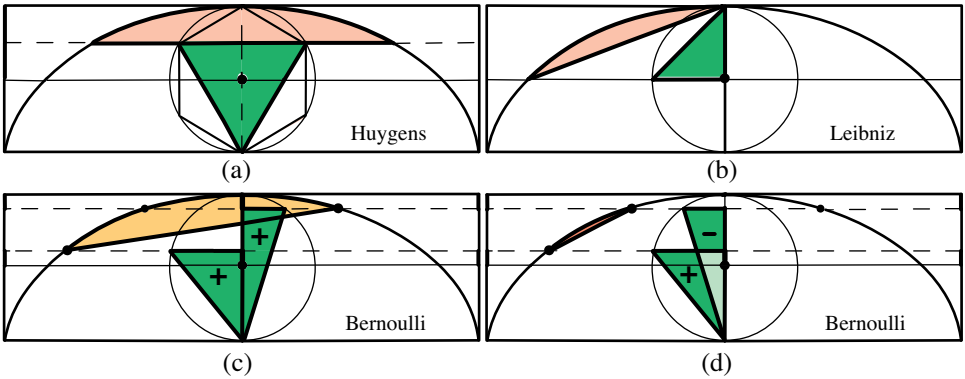


Figure 2.30: Quadrature of special cycloidal segments when the circular disk is centrally located, by (a) Huygens, (b) Leibniz, and generalizations by J. Bernoulli in (c) and (d).

In Figure 2.30b, the shaded curvilinear segment is cut off by a chord of the cycloid. In 1678, Leibniz proved that the area of this segment is equal to that of the shaded isosceles right triangle inside the disk. The quadrature results of Huygens and Leibniz are of special interest because when the disk has unit radius each curvilinear segment has the same area as a polygonal region whose area is an algebraic number, hence not a rational multiple of π. By contrast, the area of a complete cycloidal arch is 3π.

Figures 2.30c and 2.30d generalize both these results, using two horizontal dashed lines whose distances from the upper and central lines in Figures 2.30a and 2.30b are equal. In 1699, Bernoulli proved that the area of the cycloidal sector in Figure 2.30c is the sum of the areas of the two right triangles shown, and the area of the small cycloidal sector in Figure 2.30d is the difference of the areas of the triangles. Figure 2.30a is the special case in which the two dashed lines coincide, while Figure 2.30b is the special case in which the distance between the dashed lines is the radius of the disk. Bernoulli was so proud of this result that he included the diagram in Figure 2.30c on the title pages of all four volumes of his collected works. When the disk has unit radius, the areas of the Bernoulli triangles might be rational multiples of π, except for the special cases of Huygens and Leibniz, when they are not.

Encouraged by these results, we consider a segment cut from a general point of a cycloid to its highest point. In Figure 2.31a, the general point is lower than the center line, and the segmental area is the sum of the areas of the two shaded right triangles. We have verified that the area relations in Figure 2.31a can be obtained by our method using Figure 2.32, which shows a companion to Lemma 2.1:

The area of the cycloidal tail (tangent sweep) is equal to that of the adjacent circular segment (tangent cluster).

In Figure 2.31b, the general point is above the center line, and the segmental area is the difference of the triangular areas. By combining two such segments with

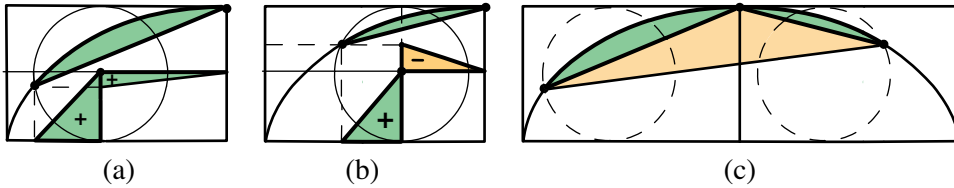


Figure 2.31: The area of a cycloidal segment is the sum of the areas of the shaded triangles in (a), and their difference in (b). Bernoulli's result in Figure 2.30c can be deduced from Figure 2.31c.

the triangle between them and using intricate dissections, we can deduce the area of a general cycloidal segment in Figure 2.31c, and in particular Bernoulli's special segments, both of whose endpoints are above the center line.

Figure 2.32b shows the cycloidal segment of Figure 2.31a inscribed in a triangle whose horizontal base has length s equal to the length of the circular arc bounding the adjacent circular segment. The result in Figure 2.31a can be deduced from the diagram in Figure 2.32c, and a similar diagram yields the result in Figure 2.31b. We omit the details.

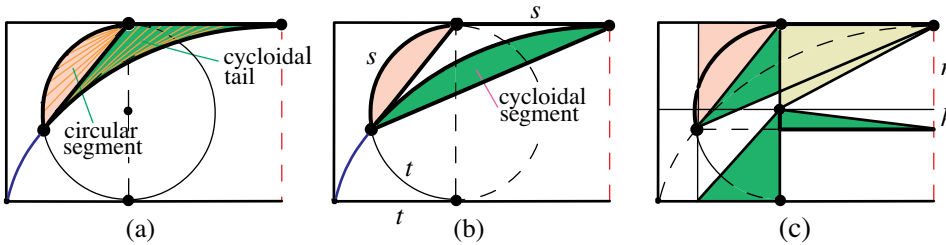


Figure 2.32: (a) Segment and tail have equal areas. (b) Triangle circumscribing the cycloidal segment. (c) Proof of the area relations in Figure 2.31a.

As a companion to the special results of Huygens, Leibniz, and Bernoulli in Figure 2.30, we include (without comment) our own special result in Figure 2.33: *the shaded curvilinear region has the same area as the adjacent square.* When the disk has unit radius the curvilinear region has area 1.

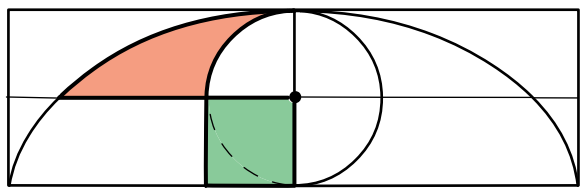


Figure 2.33: The shaded curvilinear region has the same area as the square.

2.9 EPICYCLOIDAL AND HYPOCYCLOIDAL CAPS REVISITED

This section relates the areas of epicycloidal and hypocycloidal caps with an interesting question. One circular disk of radius r rolls along the circumference of another of radius R, where the ratio of the radii is an integer, say $R = nr$. On a diameter of the rolling disk, draw a vertical vector that we use as an indicator. Question: *How many turns does the indicator make when the smaller disk rotates completely along the larger?*

Because the smaller disk returns to its original position after rolling n times along the larger disk, one would expect the number of turns of the indicator to be n, but the actual number is $n+1$ if the smaller disk rolls externally, and $n-1$ if it rolls internally.

For example, if $n = 1$ ($r = R$), the indicator makes two turns externally and no turns internally! This is illustrated in Figure 2.34a; you can try to verify it directly with two coins of the same size. Figure 2.34b illustrates what happens when $n = 2$, showing both external and internal rolling; the same occurs for general n. As the small disk rolls clockwise externally, the indicator also turns clockwise, as indicated for $n = 2$ by the intermediate positions in Figure 2.34b, and the indicator has made $n+1$ complete turns when it returns to its original position. On the other hand, when the small disk rolls internally clockwise, the indicator turns counterclockwise making $n-1$ complete turns when it returns to its original position.

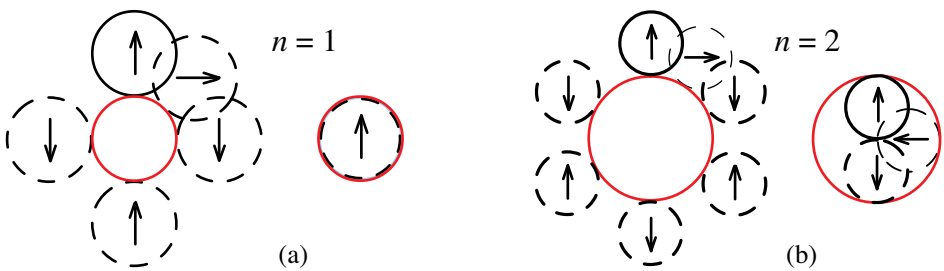


Figure 2.34: Number of turns of the indicator when disk of radius r rolls n times along the circumference of disk of radius nr. In (a), $n = 1$: the indicator makes 2 turns externally, no turns internally. In (b), $n = 2$: the indicator makes 3 turns externally, 1 turn internally.

From the fact that the number of turns of the indicator is $n \pm 1$, it can be shown that the sum of the areas of complementary epicycloidal and hypocycloidal caps is $2[D]$, where $[D]$ is the area of the rolling disk, regardless of the radii. (See Corollary 2.4.) Their difference is $4(r/R)[D]$, as obtained in Section 2.4 for their arches.

NOTES ON CHAPTER 2

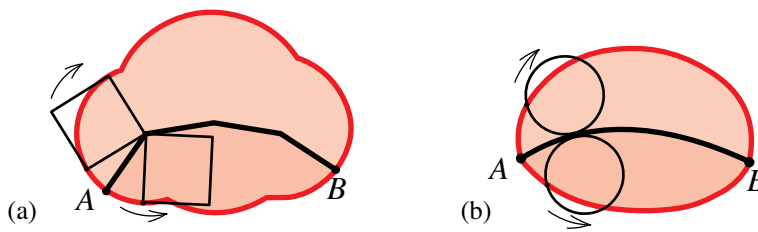
Sections 2.1 through 2.5, and 2.8, first appeared in [24], which was awarded a Lester R. Ford Award in August 2010. The material in Sections 2.6 and 2.7 has not been previously published. Some area results will be deduced differently in Chapter 3, using rolling polygons instead of circles. Arclength relations will be discussed in Chapter 3 and also in Chapter 11.

Chapter 3

CYCLOGONS AND TROCHOGONS

These problems can be easily solved by the methods developed in this chapter. The reader may wish to try solving them before reading the chapter.

Figure (a) shows a polygonal track with four edges of unit length joining A and B. A unit square rolls along one side of the track with one vertex tracing the larger arch shown, and it also rolls along the opposite side tracing the smaller arch.



Show that the sum of the arclengths of the two arches does not depend on the track, and that the same is true for the sum of the areas of the two shaded regions between the arches and the track.

Prove corresponding results for a circle rolling along a smooth curve as in (b).

CONTENTS

When a regular polygonal disk rolls along a straight line, each vertex traces a curve we call a *cyclogon*. More generally, when a regular disk with n sides rolls around another with m sides, a point rigidly attached to the rolling disk traces a curve we call a *trochogon*. This chapter determines areas of trochogonal arches without using calculus. In the limiting case when n and m tend to infinity, the polygonal disks become circular and we obtain classical results for various types of cycloidal arches described in Chapter 2.

When the tracing point is a vertex, the same elementary treatment yields arclengths of the corresponding trochogonal curves, and leads to classical results in the limiting case when the polygonal disks become circular. These arclength results are also obtained in Chapter 11 by a different method.

We also introduce *autogons*, curves traced by a point rigidly attached to an arbitrary n-gon that rolls around a fixed mirror image of itself. A comparison theorem shows that the area of the region inside an autogon is twice that of a cyclogonal arch obtained by rolling the polygon along a line. A corresponding theorem is obtained for arclengths. Limiting cases yield classical theorems of Steiner.

The results are generalized to incomplete trochogons and autogons. Applications are given to elliptic, hyperbolic, and parabolic catenaries.

3.1 INTRODUCTION

In Chapter 2, Mamikon's sweeping-tangent theorem was used to show that the area A of the cycloidal arch in Figure 3.1 is equal to three times the area C of the rolling circular disk,

$$A = 3C, \tag{3.1}$$

and the significance of the factor 3 was explained.

This chapter solves the more general problem, in which the rolling circle is replaced by a regular polygon. The problem is treated by an elementary geometrical

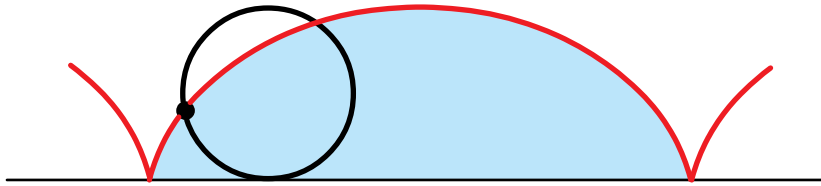


Figure 3.1: Cycloid traced by a point on the circumference of a rolling circle.

method, and the area formula for the cycloid is obtained as a limiting case. We use the formula for the area of a circular sector, but there is no need to know the cartesian or parametric equations representing the cycloid.

3.2 CYCLOGONS

A *cyclogon* is traced by a vertex of a polygon that rolls without slipping along a straight line. Like the cycloid, a cyclogon consists of a sequence of arches resting on the line, as shown by the example of a rolling regular pentagon in Figure 3.2.

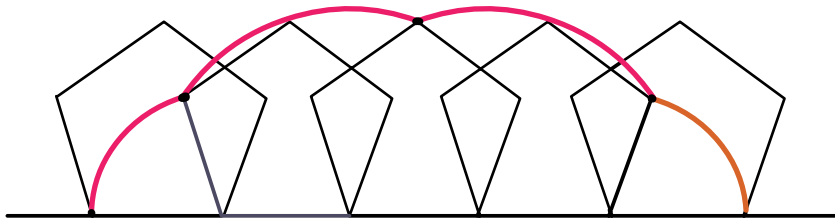


Figure 3.2: Cyclogon traced by a vertex of a rolling regular pentagon.

Each arch, in turn, is composed of circular arcs, equal in number to one fewer than the number of vertices of the polygon. The radius of each arc is the distance from the moving vertex to the pivotal vertex.

We begin with cyclogons generated by regular polygons. Let A denote the area of the region above the line and below one of the arches, let P denote the area of the rolling polygon, and let C denote the area of the disk that circumscribes the polygon. We will prove the following elegant result that generalizes (3.1).

Theorem 3.1. *For every cyclogon generated by a regular polygon*

$$A = P + 2C. \tag{3.2}$$

The circle can be regarded as the limiting case obtained by letting the number of edges of the polygon increase without bound. Similarly, the cycloid is the limiting case of a cyclogon. Then (3.1) is revealed as a limiting case of (3.2).

The proof of Theorem 3.1 for a rolling n-gon is given in Section 3.3. Before discussing the general n-gon we deduce (3.2) directly for two simple examples.

Example 1 (Rolling equilateral triangle).

Figure 3.3 shows one arch of a cyclogon traced by a vertex of a rolling equilateral triangle Δ whose edges have length a. The region under the arch and above the line consists of two equal circular sectors of radius a, and one copy of Δ. Each

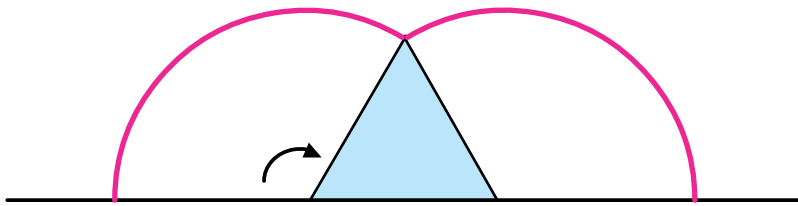


Figure 3.3: One arch of a cyclogon traced by a rolling equilateral triangle.

circular sector has area $(\pi/3)a^2$ which is also the area C of the circular disk that circumscribes Δ. Therefore

$$A = \text{area of}\,\Delta + 2 \times \frac{\pi}{3}a^2 = \text{area of}\,\Delta\, + \,2C,$$

which proves (3.2) in this case.

Example 2 (Rolling square).

Figure 3.4 shows a cyclogon traced by a vertex of a rolling square. The region under the arch consists of two right triangles plus three circular quadrants, two of radius a (the edge-length of the square), and one of radius $a\sqrt{2}$ (the diagonal of the

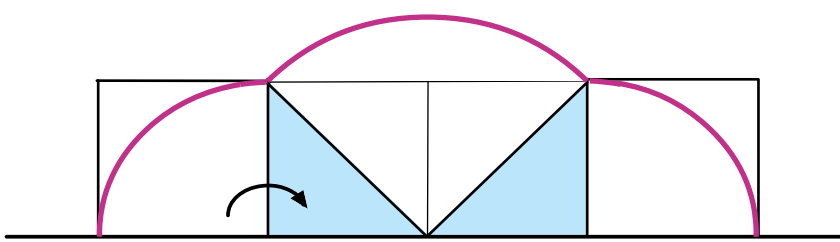


Figure 3.4: One arch of a cyclogon traced by a rolling square.

square). The two right triangles have total area a^2, the area of the rolling square, and the total area of the three circular quadrants is

$$2 \times \frac{\pi}{4}a^2 + \frac{\pi}{4}(a\sqrt{2})^2 = 2 \times \pi(a\,\frac{\sqrt{2}}{2})^2 = 2C.$$

Therefore we have $A = a^2 + 2C$, which proves (3.2) in this case as well.

3.3 AREA OF A CYCLOGONAL ARCH GENERATED BY A REGULAR POLYGON

In the general case of a regular polygon with n vertices, the region under one arch of the cyclogon consists of $n-2$ triangles and $n-1$ circular sectors, each subtending an angle of $2\pi/n$ radians. These triangles can be regarded as footprints left by the triangular pieces obtained by dissecting the original polygon with diagonals from a vertex to each of the nonadjacent vertices, as illustrated in Figure 3.5 for $n = 6$.

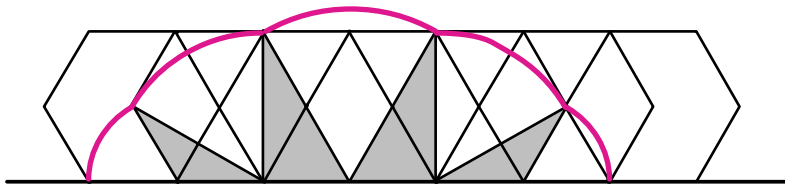


Figure 3.5: Footprints left by triangular pieces of a rolling hexagon.

The sum of the areas of the triangles is equal to the area P of the region enclosed by the regular polygon. This is illustrated for the regular hexagon in Figure 3.6.

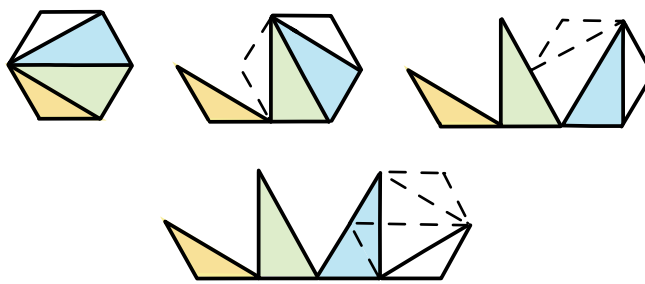


Figure 3.6: Visualizing the distribution of the footprints in Figure 3.5.

The radii of the circular sectors are the lengths of the segments from one vertex to the remaining $n-1$ vertices. A sector of radius r_k subtending an angle of $2\pi/n$ radians has area $\pi r_k^2/n$, so the sum of the areas of the $n-1$ sectors is equal to

$$\frac{\pi}{n}\sum_{k=1}^{n-1} r_k^2.$$

From Theorem 14.2 in Chapter 14, it follows that the sum of the squares of the radii is given by

$$\sum_{k=1}^{n-1} r_k^2 = 2nr^2, \tag{3.3}$$

where r is the radius of the circle that circumscribes the polygon. Therefore the sum of the areas of the sectors is equal to $2\pi r^2$, which is $2C$, twice the area of the circumscribing disk, thus proving $A = P + 2C$, as stated in (3.2) of Theorem 3.1.

Another interpretation of Theorem 3.1.

When (3.2) of Theorem 3.1 is written as

$$A - P = 2C$$

it reveals an interesting fact. Consider all regular polygons circumscribed by a common disk of area C. As we change the polygon, areas A and P will change, but their difference $A - P$ does not change because it is always equal to $2C$.

Some examples are shown in Figure 3.7, where each shaded region has area $A - P$. It is not obvious from the figure that all the shaded regions have equal area. Nevertheless, Theorem 3.1 tells us that each area is $2C$. The last example is the degenerate case when the polygon consists of two edges, and the first example is the limiting case when the regular polygon becomes a circle and the cyclogon becomes a cycloid.

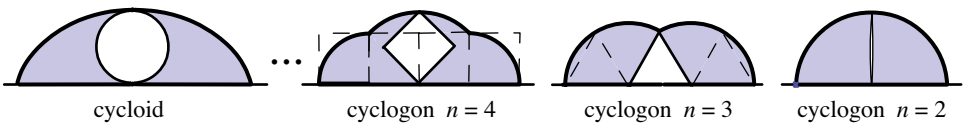


Figure 3.7: All shaded regions have equal area.

3.4 TROCHOGONS: GENERALIZED CYCLOGONS

As mentioned earlier, a cycloid is the curve traced by a point on the circumference of a circular disk that rolls without slipping along a straight line. It consists of a periodic sequence of congruent arches resting on the line. If the point is rigidly attached to the disk but not on the circumference it traces a *curtate cycloid* if the tracing point lies inside the disk, and a *prolate cycloid* if it lies outside the disk. Figure 3.8 shows an example of each type.

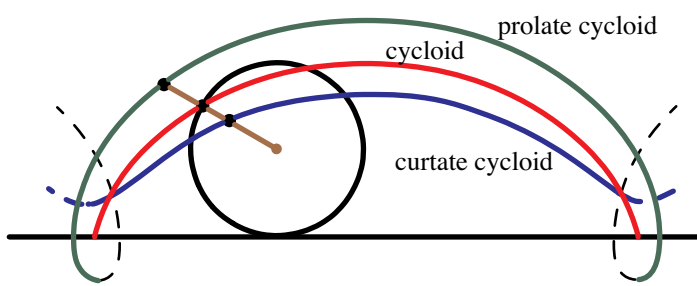


Figure 3.8: A cycloid, a curtate cycloid, and a prolate cycloid traced by a point rigidly attached to a rolling disk.

If the rolling disk is replaced by a regular polygon, each vertex traces a cyclogon. In Theorem 3.1 we proved that the area A of the region under one arch of a cyclogon

is given by

$$A = P + 2C,$$

where P is the area of the rolling polygon and C is the area of the disk that circumscribes the polygon.

To better understand why the term $2C$ appears in this formula, we treat the more general area problem for curtate and prolate cyclogons and obtain a result that is surprisingly simple:

Theorem 3.2. *If C_z is the area of a disk whose radius is the distance from the center of the rolling polygon to the tracing point z, then*

$$A = P + C + C_z. \tag{3.4}$$

When z is on the circumference of the rolling disk, we have $C_z = C$ and we get (3.2). In the limiting case when P approaches C, this gives a known result, $A = 2C + C_z$. We will deduce (3.4) as a limiting case of a more general result concerning trochogons, or generalized cyclogons, the main topic of this section.

Here we consider a more general situation in which a curve is traced by a point z on a regular polygonal disk with n sides rolling around another regular polygonal disk with m sides. All edges of the two regular polygons are assumed to have the same length. A point z attached rigidly to the n-gon traces an arch consisting of n circular arcs before repeating the pattern periodically. We call this curve a *trochogon*. It is called an *epitrochogon* if the n-gon rolls outside the m-gon, and a *hypotrochogon* if it rolls inside the m-gon. The trochogon is curtate if z is inside the n-gon, and prolate (with loops) if z is outside the n-gon. If z is at a vertex it traces an *epicyclogon* or a *hypocyclogon*. Figure 3.9 shows a curtate epitrochogon obtained by rolling a square ($n = 4$) outside a 24-gon ($m = 24$).

Area of a trochogonal arch.

The main result of this section is a simple and elegant formula for the area A of the region between a general trochogonal arch and the fixed polygon. We call this the area of the trochogonal arch. It is given by (3.5) in Theorem 3.3, in which P_n denotes the area of the regular n-gon that rolls around a regular m-gon having edges of the same length, C is the area of the disk that circumscribes the n-gon, and C_z is the area of the concentric disk whose boundary passes through the tracing point z.

Theorem 3.3. *The area A of the trochogonal arch is given by*

$$A = P_n + (1 \pm \frac{n}{m})(C_z + C), \tag{3.5}$$

with the plus sign for an epitrochogon and the minus sign for a hypotrochogon.

Proof. The following elementary proof of (3.5) does not use integral calculus or Mamikon's sweeping-tangent theorem. It is illustrated in Figure 3.9, with $n = 4$ and $m = 24$, which displays the essential features required for treating a general

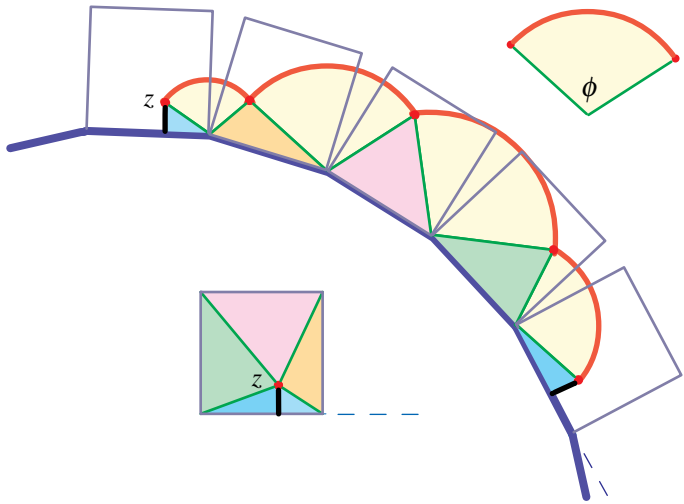


Figure 3.9: A curtate epitrochogonal arch traced by a point inside a square rolling outside a regular 24-gon.

regular n-gon rolling outside a regular m-gon. The tracing point z is inside the square, and the arch it generates consists of four circular sectors and five triangles, shown shaded. The lower portion of Figure 3.9 shows how the five triangular pieces fill the square. Because of periodicity, the first and last right triangles outside the 24-gon together have the same area as the bottom triangle in the square. So A is equal to P_4, the area of the rolling square, plus the sum of the areas of the four circular sectors.

In the general case of a regular n-gon rolling outside a regular m-gon, the tracing point z attached to the n-gon generates an arch consisting of n circular sectors together with a set of triangles that provide a dissection of the n-gon. So the area A of any trochogonal arch is equal to P_n, the area of the rolling n-gon, plus the areas of n circular sectors, the kth sector having area $\phi r_k^2/2$ where ϕ is the common angle (in radians) subtended by each sector and $r_1, \ldots, r_n$ are the radii of the sectors. Radius r_k is the distance from the tracing point z to the kth vertex of the rolling polygon. Thus, we have

$$A = P_n + \frac{1}{2}\phi \sum_{k=1}^{n} r_k^2. \tag{3.6}$$

It is easy to see that $\phi = 2\pi/n + 2\pi/m$, the sum of two exterior angles, so (3.6) becomes

$$A = P_n + (1 + \frac{n}{m})\,\frac{\pi}{n} \sum_{k=1}^{n} r_k^2. \tag{3.7}$$

To evaluate the sum of squares in (3.7) we invoke (14.3) from Chapter 14, where r is the radius of the circle circumscribing the rolling n-gon, and $|z|$ is the distance

from the center of the n-gon to z, and we find

$$\frac{\pi}{n}\sum_{k=1}^{n} r_k^2 = \pi|z|^2 + \pi r^2 = C_z + C.$$

Use this in (3.7) to get a formula for the area of an epitrochogonal arch:

$$A = P_n + (1 + \frac{n}{m})(C_z + C). \tag{3.8}$$

Incidentally, if the rolling n-gon rolls inside the m-gon, the same analysis shows that the area of a hypotrochogonal arch is

$$A = P_n + (1 - \frac{n}{m})(C_z + C), \tag{3.9}$$

so (3.8) and (3.9) can be combined to give (3.5). When $m \to \infty$ this gives (3.4).

Limiting case: epitrochoid and hypotrochoid.

Area relations for epitrochoids and hypotrochoids were obtained in Chapter 2. For example, Theorem 2.2 implies that the area of a full trochoidal arch obtained by rolling a disk of radius r and area C once around a fixed disk of radius R is equal to $(3 \pm 2r/R)C$, with the $+$ sign for the epitrochoid and the $-$ sign for the hypotrochoid. This can also be deduced as a limiting case of Theorem 3.3 if we let both n and m tend to ∞ in such a way that their ratio $n/m \to r/R$. This limiting case of (3.5) also generalizes Theorem 2.2 for a complete epitrochoidal and hypotrochoidal arch traced by a general point z of the rolling disk. In the notation of Theorem 3.1, the polygonal area P becomes C and we obtain:

Corollary 3.1. *The area A of a trochoidal arch is given by*

$$A = C + (1 \pm \frac{r}{R})(C_z + C), \tag{3.10}$$

where the $+$ sign is used for the epitrochoid and the $-$ sign for the hypotrochoid.

When the tracing point z is on the boundary of the rolling disk, then $C_z = C$ and (3.10) reduces to $A = (3 \pm 2r/R)C$, as implied by Theorem 2.2.

Dependence on the tracing point.

Each row of Figure 3.10 shows two trochogons traced by two different points z at the same distance from the center of the rolling polygon. In Figure 3.10a, where the rolling polygon is a 2-gon, the first arch is traced by a point z outside the 2-gon at distance from the center equal to half the edge length of the 2-gon, and the second is traced by a point z at a vertex. In both cases, the area C_z is the same so by (3.4) the area A is the same. In Figure 3.10b, where the rolling polygon is an equilateral triangle, the first trochogon is traced by a midpoint of the base of the triangle, and the second (curtate) trochogon is traced by a point z inside the triangle at the same

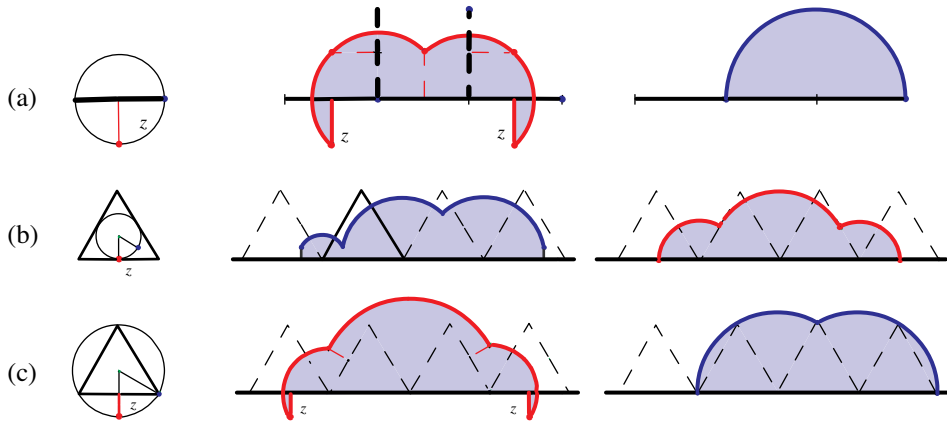


Figure 3.10: In each row, the shaded regions have equal area.

distance from the center as the midpoint of the base. Consequently, C_z is the same for both trochogons, so the areas A of the arches are also the same.

In Figure 3.10c the rolling triangle is that of Figure 3.10b. The first trochogon is a prolate trochogon traced by a point z outside the triangle at the same distance from the center of the triangle as a vertex. The second trochogon is traced by a vertex. The two trochogons in this case have equal areas.

3.5 SPECIAL TROCHOGONS

Tracing point at a vertex.

Return now to (3.5) and take the tracing point z at a vertex of the rolling n-gon. Then the areas C_z and C of the disks are equal, and (3.5) gives the area of one arch of an epicyclogon or hypocyclogon:

$$A = P_n + 2(1 \pm \frac{n}{m})C. \tag{3.11}$$

In the limiting case when both n and m tend to ∞ in such a way that $n/m \to r/R$, we find $P \to C$ and (3.11) gives us a known result for the area of one arch of the classical epicycloid or hypocycloid:

$$A = (3 \pm 2\frac{r}{R})C. \tag{3.12}$$

A special case of (3.11) is the area for the *cardiogon* (Figure 3.11), which is an epicyclogon with $n/m = 1$:

$$A = P_n + 4C. \tag{3.13}$$

When $n \to \infty$, then $P_n \to C$, the tracing curve becomes a cardioid, and (3.13) gives us $A = 5C$. This implies a classical result that the area of the region bounded

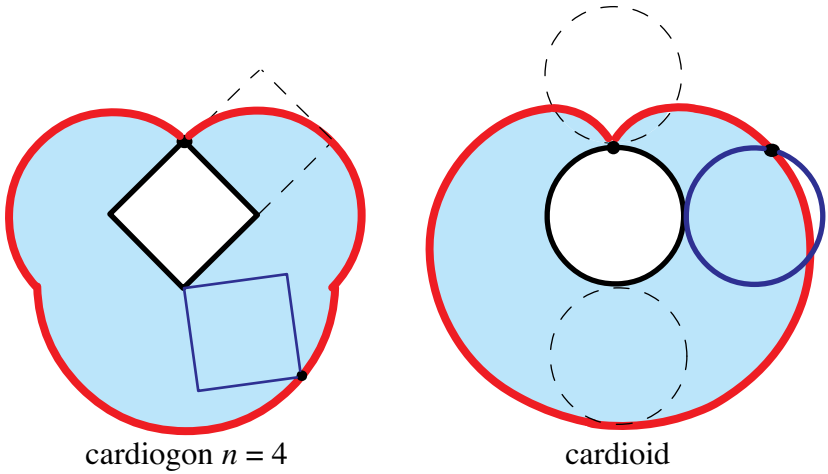


Figure 3.11: A cardiogon traced by the vertex of an n-gon rolling outside an n-gon. The cardiogon becomes a cardioid as $n \to \infty$.

by a cardioid is equal to $6C$, because the cardioidal arch, of area $5C$, together with the inner disk of area C, fill the cardioid with area $6C$.

Another special case is the *nephrogon* (Figure 3.12), an epicyclogon with $n/m = 1/2$, for which (3.11) gives

$$A = P_n + 3C. \tag{3.14}$$

When $n \to \infty$, (3.14) becomes $A = 4C$, for the area of one arch of a nephroid. The nephroid itself encloses two such arches, each of area $4C$, plus the inner disk of area $4C$, giving another proof of the known result that a nephroid encloses a region whose area is $12C$.

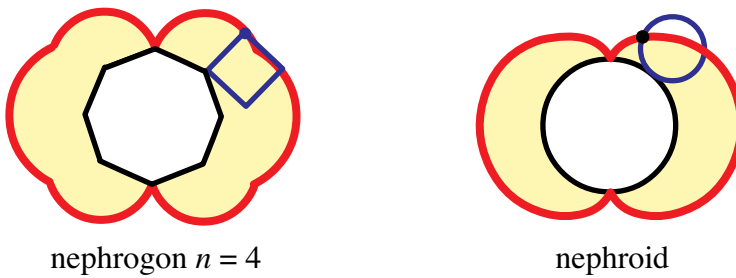


Figure 3.12: A nephrogon traced by the vertex of an n-gon rolling outside a $2n$-gon. The nephrogon becomes a nephroid as $n \to \infty$.

A related figure is the *astrogon* (Figure 3.13), a hypocyclogon with $n/m = 1/4$, for which (3.11) yields

$$A = P_n + \frac{3}{2}C. \tag{3.15}$$

When $n \to \infty$, (3.15) gives $A = (5/2)C$ for an astroid, which is a hypocycloid with

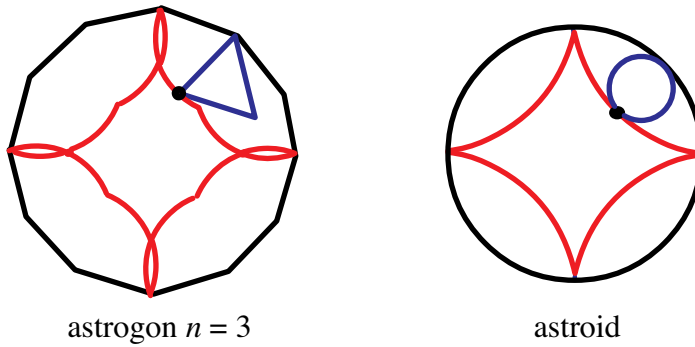


Figure 3.13: An astrogon traced by the vertex of an n-gon rolling inside a $4n$-gon. The astrogon becomes an astroid as $n \to \infty$.

four cusps ($r/R = 1/4$). The four arches between the hypocycloid and the outer circle (of area $16C$) have a total area of $4A = 10C$, so the region inside the astroid has area $6C$, another classical result obtained without calculus.

Another special case of interest is the *deltogon* (Figure 3.14), a hypocyclogon with $n/m = 1/3$, for which (3.11) gives

$$A = P_n + \frac{4}{3}C. \tag{3.16}$$

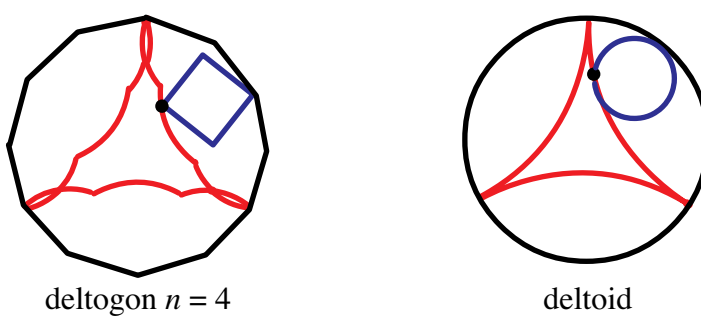


Figure 3.14: A deltogon traced by the vertex of an n-gon rolling inside a $3n$-gon. The deltogon becomes a deltoid as $n \to \infty$.

When $n \to \infty$, (3.16) becomes $A = (7/3)C$ for the deltoid, which is a hypocycloid with three cusps ($r/R = 1/3$). The three arches between the deltoid and the fixed circle have a total area of $3A = 7C$, the fixed circle has area $9C$, so the region inside the deltoid has area $2C$, another known result.

A somewhat suprising example is what we call a *diamogon*, a hypocyclogon with $n/m = 1/2$. The curve is traced by a point z at a vertex of an n-gon rolling inside

a $2n$-gon. When the n-gon makes one circuit around the inside of the $2n$-gon, it traces two curves each consisting of $n-1$ circular arcs situated symmetrically about a diameter of the $2n$-gon. Examples with $n = 3$ and 4 are shown in Figure 3.15.

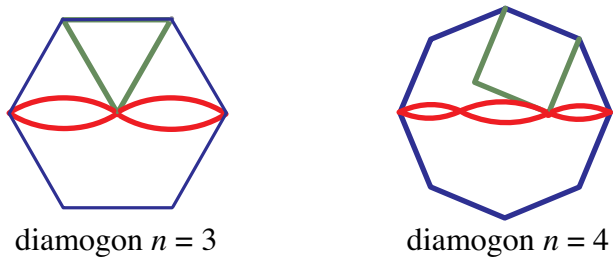


Figure 3.15: Diamogons traced by a vertex of an n-gon rolling inside a $2n$-gon.

Using (3.9) we find that the area of one arch of the diamogon is $A = P_n + C$ because $C_z = C$. The two arches between the diamogon and the outer polygon have area $2A = 2(P_n + C)$. In the limiting case when $n \to \infty$ this becomes $2A = 4C$. But $4C$ is the area of the fixed circular disk, which means that the area of the region common to the two diamogons tends to zero. In other words, when $n \to \infty$ the diamogon turns into a diameter of the fixed circle, traced twice.

Another geometric interpretation of Theorem 3.3.

When (3.5) of Theorem 3.3 is written in the form

$$A - P_n = (1 \pm \frac{n}{m})(C_z + C)$$

it tells us that the difference $A - P_n$ will be the same for all regular n-gons as long as the right member of the equation is constant. The area C will be constant if all the n-gons are inscribed in the same circle, the area C_z will not change if the tracing point z is chosen at a fixed distance from the center of the rolling polygon, and the factor multiplying $(C_z + C)$ will not change if the ratio n/m is constant. Figure 3.16 shows examples with $n/m = 1$, $C_z = C$, and $A - P_n = 4C$.

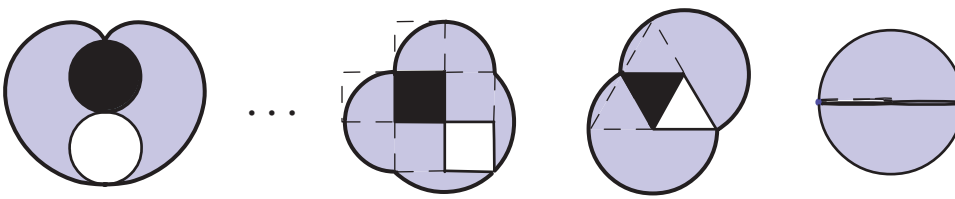


Figure 3.16: All lightly shaded regions have equal area.

Tracing point not at a vertex.

Now we consider an example of a hypotrochogon traced by a point z not at a vertex of the n-gon. We take $n/m = 1/2$ and call the hypotrochogon an *ellipsogon* because the limiting case $n \to \infty$ gives an ellipse. Figure 3.17 shows an example of a square rolling inside an octagon with the tracing point z inside the square. In this case the ellipsogon traces two arches, each consisting of four circular arcs.

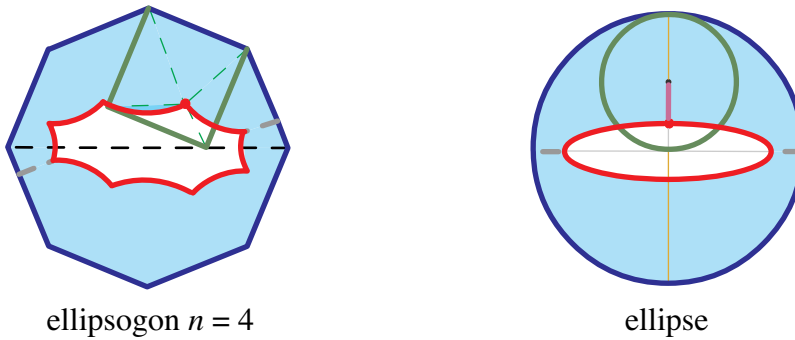


Figure 3.17: An ellipsogon traced by a point inside an n-gon rolling inside a $2n$-gon. The ellipsogon becomes an ellipse as $n \to \infty$.

In the limiting case $n \to \infty$, (3.9) shows that the area of one arch is given by $A = C + (C_z + C)/2$, so the two arches fill out a region of area $2A = 3C + C_z$. The limiting configuration of the ellipsogon is an ellipse enclosing an area equal to $4C - 2A = C - C_z$. If the radius of the inner circle is r and if the distance from z to the center of the inner circle is s, then

$$C - C_z = \pi(r^2 - s^2) = \pi(r + s)(r - s).$$

The distances $r + s$ and $r - s$ are the lengths of the semiaxes $a = r + s$ and $b = r - s$ of the ellipse, so we get $C - C_z = \pi ab$, the usual formula for the area of an ellipse.

The point z also traces an ellipsogon if it is outside the rolling n-gon. If the point z is inside or outside the rolling n-gon and then moves toward a vertex, the ellipsogon becomes a diamogon which, in turn, becomes a diameter as $n \to \infty$.

3.6 ARCLENGTH OF CYCLOGONAL ARCHES

Areas of cyclogonal arches generated by regular polygons were treated in Theorems 3.1 and 3.2. We turn next to the problem of calculating arclengths. To better appreciate the essence of the problem we treat both the area and arclength for arches generated by a general convex n-gon, not necessarily regular. First we consider a general convex n-gon rolling along a line with the tracing point z inside the polygon, as illustrated by the example in Figure 3.18.

General convex n-gon rolling along a line; tracing point inside.

The rolling quadrilateral in Figure 3.18 displays all the essential features required for treating a general convex n-gon.

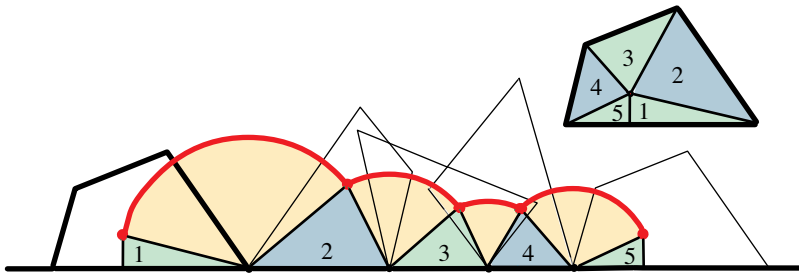


Figure 3.18: A curtate cyclogon traced by a point inside a quadrilateral rolling along a line.

In one revolution the tracing point traces four circular arcs. The cyclogonal region below the arcs and above the line consists of four circular sectors and a set of triangles that form a dissection of the quadrilateral as indicated in Figure 3.18. For a general n-gon making one revolution along a line, the tracing point z generates a cyclogonal region consisting of n circular sectors together with a set of triangles that provide a dissection of the n-gon. We wish to determine the area of the cyclogonal region swept out and the length of the corresponding arch.

The area A of the region is equal to the area P_n of the rolling n-gon, plus the sum of the areas of n circular sectors. The kth sector has area $(1/2)\phi_k r_k^2$, where $r_1, \ldots, r_n$ are the radii of the sectors, and ϕ_k is the angle (in radians) subtended by the sector of radius r_k. Note that r_k is the distance from the tracing point z to the kth vertex of the rolling polygon, and ϕ_k is the exterior angle of the polygon at that vertex. Thus we have

$$A = P_n + \frac{1}{2}\sum_{k=1}^{n} \phi_k r_k^2. \tag{3.17}$$

The length of the circular arc subtended by sector k is $\phi_k r_k$. The sum of the lengths is the length L of one arch of the curtate cyclogon:

$$L = \sum_{k=1}^{n} \phi_k r_k. \tag{3.18}$$

The sums on the right of (3.17) and (3.18) cannot be simplified until more is known about the radii and the exterior angles. Therefore we turn next to regular rolling polygons.

Regular n-gon rolling along a line.

When the rolling n-gon is regular, it generates a cyclogon if the tracing point is at a vertex, a curtate cyclogon if the tracing point is inside the polygon, and a prolate

cyclogon if it is outside. In this case each exterior angle ϕ_k is equal to $2\pi/n$, and the respective formulas (3.17) and (3.18) become

$$A = P_n + \frac{\pi}{n}\sum_{k=1}^{n} r_k^2, \tag{3.19}$$

and

$$L = \frac{2\pi}{n}\sum_{k=1}^{n} r_k. \tag{3.20}$$

When the tracing point is at a vertex the sum in (3.19) can be simplified with the help of (14.3) in Chapter 14 which, as we have already seen, leads to Theorems 3.1 and 3.2.

There is no analog of (14.3) that can be used in (3.20) to obtain a simple arclength formula. This is not surprising because calculating arclength is usually more difficult than calculating area. For example, the area of the region enclosed by an ellipse can be easily determined (with or without calculus), but calculating the arclength of a general ellipse requires elliptic integrals. We now determine (without using calculus) the arclengths of epicyclogonal and hypocyclogonal arches when the tracing point z is at a vertex of the rolling regular polygon.

Arclength of a cyclogon (tracing point at a vertex).

When the tracing point z is at a vertex of the regular n-gon, the trochogon is a cyclogon, and (3.19) becomes (3.2) of Theorem 3.1. As already noted, in the limiting case when $n \to \infty$, P_n becomes C, the cyclogon becomes a cycloid, and (3.2) gives us the classical result $A = 3C$.

Another classical result states that the length L of a cycloidal arch is $4D$, where D is the diameter of the rolling circle. The next theorem generalizes this result.

Theorem 3.4. *The arclength of a cyclogonal arch traced by a vertex of a rolling regular n-gon is given by*

$$L = 4D\,(\frac{\pi}{2n}\cot\frac{\pi}{2n}), \tag{3.21}$$

where D is the diameter of the circle that circumscribes the n-gon.

This implies that $L \to 4D$ as $n \to \infty$ because

$$\lim_{n\to\infty}(\frac{\pi}{2n}\cot\frac{\pi}{2n}) = 1.$$

The quantity

$$\frac{\pi}{2n}\cot\frac{\pi}{2n}$$

is surprisingly close to 1, even for small values of n. For example, for $n = 3, 4, 5, 6$ its value to two decimals is 0.91, 0.95, 0.97, 0.98.

Proof of Theorem 3.4. Start with (3.20) and note that r_k is the distance from the tracing point to the kth vertex of the polygon. When the tracing point is one of the vertices, (3.20) becomes

$$L = \frac{2\pi}{n} \sum_{k=1}^{n-1} r_k, \tag{3.22}$$

where now $r_1, \ldots, r_{n-1}$ are the lengths of the segments from one vertex to each of the remaining $n-1$ vertices.

Figure 3.19 shows two examples, (a) a regular pentagon, and (b) a regular hexagon. In (a) there are four segments r_1, r_2, r_3, r_4 symmetrically located about a diameter of length D. In (b) there are five segments, one of which is a diameter, the other four being located symmetrically about it.

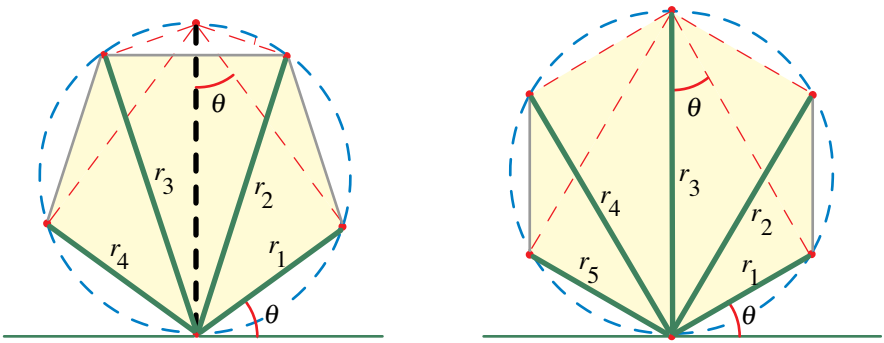


Figure 3.19: Segments drawn from one vertex of a regular polygon to the remaining vertices.

In a general regular n-gon, each segment r_k together with a diameter D determines a right triangle inscribed in a semicircle with the diameter as hypotenuse. One of the acute angles of the right triangle is $k\theta = k\pi/n$, so

$$r_k = D \sin \frac{k\pi}{n},$$

and (3.22) becomes

$$L = \frac{2\pi D}{n} \sum_{k=1}^{n-1} \sin \frac{k\pi}{n}. \tag{3.23}$$

Fortunately there is a trigonometric identity that is tailor-made to evaluate the sum of sines in (3.23). It states that for any real x not an integer multiple of π we have

$$\sum_{k=1}^{n} \sin(2kx) = \frac{\sin(n+1)x \, \sin nx}{\sin x}. \tag{3.24}$$

This identity is a disguised form of an elementary formula for the sum of a geometric progression:

$$\sum_{k=1}^{n} z^k = z\,\frac{z^n - 1}{z - 1},$$

valid for any complex $z \neq 1$. When we take $z = e^{2ix}$ in this relation we get

$$\sum_{k=1}^{n} e^{2ikx} = e^{2ix}\frac{e^{2inx} - 1}{e^{2ix} - 1} = e^{2ix}\frac{e^{inx}(e^{inx} - e^{-inx})}{e^{ix}(e^{ix} - e^{-ix})} = e^{i(n+1)x}\frac{\sin nx}{\sin x}. \tag{3.25}$$

Now equate imaginary parts of this last equation to obtain (3.24). When $x = \pi/(2n)$, the term with $k = n$ in (3.24) vanishes and we find

$$\sum_{k=1}^{n-1} \sin\frac{k\pi}{n} = \frac{\sin(\frac{\pi}{2n} + \frac{\pi}{2})\sin(\frac{\pi}{2})}{\sin\frac{\pi}{2n}} = \cot\frac{\pi}{2n}.$$

Thus we see that (3.23) becomes (3.21), which gives the arclength of every cyclogon.

If we equate real parts of (3.25) we also obtain a cosine formula:

$$\sum_{k=1}^{n} \cos(2kx) = \frac{\cos(n+1)x\,\sin nx}{\sin x}.$$

Replacing $\cos(2kx)$ by $1 - 2\sin^2(kx)$ gives

$$2\sum_{k=1}^{n} \sin^2(kx) = n - \frac{\cos(n+1)x\,\sin nx}{\sin x}, \tag{3.26}$$

a result we will use in Section 3.9.

3.7 ARCLENGTH OF EPICYCLOGONS AND HYPOCYCLOGONS

We turn now to the more general case of a regular n-gon rolling along a regular m-gon, where the edges of the two polygons have equal length. The corresponding curves traced by a vertex of the n-gon are called epicyclogons and hypocyclogons. The next theorem reduces their arclengths to that of a cyclogon (the same n-gon rolling on a straight line). The cyclogonal arclength is given by (3.21) and is denoted here by L_o.

Theorem 3.5. *The length L of one arch of an epicyclogon or a hypocyclogon traced by the vertex of a regular polygon is related to the corresponding cyclogonal arclength L_o by*

$$L = (1 \pm \frac{n}{m})L_o, \tag{3.27}$$

where the $+$ sign is used for the epicyclogon and the $-$ sign for the hypocyclogon.

Proof. The analysis used in deriving (3.20) for a polygon rolling along a line can also be applied to this more general case. The only change is that in (3.18) the angle ϕ_k through which the polygon rolls about the kth vertex is equal to $2\pi/n + 2\pi/m$ for the epicyclogon, and $2\pi/n - 2\pi/m$ for the hypocyclogon. For the length of one arch we have, instead of (3.20), the relation

$$L = (\frac{2\pi}{n} \pm \frac{2\pi}{m}) \sum_{k=1}^{n} r_k \,, \tag{3.28}$$

where the plus sign is used for the epicyclogon and the minus sign for the hypocyclogon. Once again we have $r_k = D\sin(k\pi/n)$, where D is the diameter of the circle that circumscribes the rolling polygon, and instead of (3.21) we obtain (3.27) of Theorem 3.5. As expected, (3.21) is the limiting case of (3.27) when $m \to \infty$.

We can obtain the limiting case of a circle of radius r rolling around a fixed circle of radius R if we let both n and m tend to infinity in such a way that their ratio $n/m \to r/R$. Then the limiting value of (3.27) gives the lengths of the corresponding epicycloid and hypocycloid. If we note that the diameter D of the rolling circle is $2r$, we find the limiting value of the arclength is

$$L = 4D(1 \pm \frac{D}{2R}) = 8r(1 \pm \frac{r}{R}).$$

3.8 SOME SPECIAL TROCHOGONS

Some classic trochoids have names suggested by their shapes, for example, cardioid, nephroid, astroid, and deltoid. We have given similar names to the trochogons having these trochoids as limits. For example, a cardiogon is generated by a regular n-gon rolling around a fixed copy of itself. Figure 3.20 shows an example with $n = 6$. As $n \to \infty$ the cardiogon becomes a cardioid.

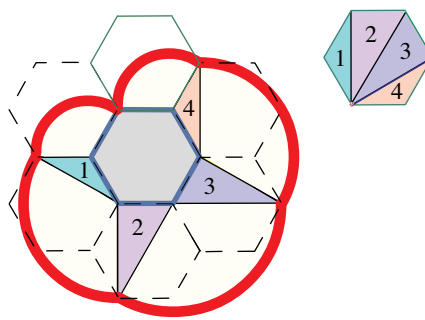


Figure 3.20: A cardiogon generated by a regular hexagon rolling around a copy of itself.

Table 1 gives the areas of one arch of a few special trochogons as derived from (3.11), together with known classical results for the corresponding trochoids. The trochoidal arclengths are also listed. We recall that P_n is the area of the rolling regular n-gon, C is the area of the circumscribing circle, and D is its diameter.

Trochogon	**Factor** $1\pm n/m$	**Area**	**Trochoid**	**Area**	**Arclength**
cyclogon	$1+n/\infty = 1$	$P_n + 2C$	cycloid	$3C$	$4D$
cardiogon	$1 + n/m = 2$	$P_n + 4C$	cardioid	$5C$	$8D$
nephrogon	$1+n/m = 3/2$	$P_n + 3C$	nephroid	$4C$	$6D$
astrogon	$1-n/m = 3/4$	$P_n + 3C/2$	astroid	$5C/2$	$3D$
deltogon	$1-n/m = 2/3$	$P_n + 4C/3$	deltoid	$7C/3$	$8D/3$
diamogon	$1-n/m = 1/2$	$P_n + C$	diameter	$2C$	$2D$

Table 1. Areas of one arch of some special trochogons, and areas and arclengths for the corresponding trochoids. Arclengths of trochogons are given by (3.27).

The last entry may seem somewhat surprising. What we call a diamogon is a hypocyclogon with $n/m = 1/2$. The curve is traced by a point z at a vertex of an n-gon rolling inside a $2n$-gon. When the n-gon makes one circuit around the inside of the $2n$-gon, it traces out two curves, each consisting of $n-1$ circular arcs situated symmetrically about a diameter of the $2n$-gon. Using (3.27) we find that the length L_{diam} of one arch of the diamogon is

$$L_{\text{diam}} = 2D(\frac{\pi}{2n}\cot\frac{\pi}{2n}).$$

When $n \to \infty$ this tends to $2D$, the diameter of the fixed circle. This is consistent with the fact that when $n \to \infty$ one arch of the diamogon turns into a diameter of the fixed circle. The region between the arch and the fixed circle is a semicircular disk of area $2C$, half the area of the fixed circle.

The history of many classic curves can be found on the web site

http://www-groups.dcs.st-andrews.ac.uk/history/Curves,

where one can also see animation related to these curves. Another rich source is Yates [70], which also contains formulas for areas and arclengths.

3.9 INCOMPLETE TROCHOGONS

In Chapter 2 we found cycloidal area relations when the rolling disk makes an incomplete rotation along a line or, more generally, around a circular disk. This section finds corresponding area and arclength formulas for incomplete trochogons traced by a vertex of a rolling polygon that makes an incomplete rotation along a line or a polygon.

Incomplete cyclogons.

Start with a cyclogon traced by a vertex of regular n-gon rolling a line, as in Section 3.6, but rotate the polygon through only the first p of its n edges, where $p \le n$. Then the formulas for area A_o^p and arclength L_o^p replacing (3.19) and (3.20) are

$$A_o^p = P_n^p + \frac{\pi}{n}\sum_{k=1}^{p} r_k^2,$$

$$L_o^p = \frac{2\pi}{n}\sum_{k=1}^{p} r_k,$$

where the superscript p reminds us that the polygon has rolled through p edges, and the subscript o reminds us that the polygon rolls along a line. Area P_n^p is that of a segment of the rolling n-gon consisting of adjacent triangular footprints like those in Figure 3.6. As before,

$$r_k = D\sin\frac{k\pi}{n},$$

where D is the diameter of the circle that circumscribes the rolling n-gon. For this value of r_k, the area and arclength formulas become

$$A_o^p = P_n^p + \frac{\pi D^2}{n}\sum_{k=1}^{p}\sin^2\frac{k\pi}{n}, \tag{3.29}$$

$$L_o^p = \frac{2\pi D}{n}\sum_{k=1}^{p}\sin\frac{k\pi}{n}. \tag{3.30}$$

Using (3.26) with $n = p$ to evaluate (3.29), we find:

Theorem 3.6. *The area of a cyclogonal sector is given by*

$$A_o^p = P_n^p + \frac{\pi D^2}{2n}[p - \frac{\sin\frac{p\pi}{n}}{\sin\frac{\pi}{n}}\cos(p+1)\frac{\pi}{n}]. \tag{3.31}$$

Similarly, using (3.24) with $n = p$ to evaluate (3.30), we obtain:

Theorem 3.7. *The arclength of an incomplete cyclogon is given by*

$$L_o^p = \frac{2\pi D}{n}\frac{\sin\frac{p\pi}{2n}}{\sin\frac{\pi}{2n}}\sin(p+1)\frac{\pi}{2n}. \tag{3.32}$$

Let $\omega = 2\pi p/n$. This is the angle swept by the radius of the rolling polygonal disk when it rolls through p of its edges. In terms of ω, (3.31) and (3.32) become

$$A_o^p = P_n^p + \frac{D^2}{4}\omega - \frac{D^2}{2}\frac{\frac{\pi}{n}}{\sin\frac{\pi}{n}}\sin\frac{\omega}{2}\cos(\frac{\omega}{2}+\frac{\pi}{n}), \tag{3.33}$$

$$L_o^p = 4D\frac{\pi}{2n}\frac{1}{\sin\frac{\pi}{2n}}\sin\frac{\omega}{4}\sin(\frac{\omega}{4}+\frac{\pi}{2n}). \tag{3.34}$$

Now let n and p tend to ∞ in such a way that the ratio p/n maintains the constant value $\omega/(2\pi)$. Denote the limiting values of A_o^p and L_o^p by A_o^ω and L_o^ω. It is easily seen that $P_n^p \to C^\omega$ where C^ω is the area of a circular segment, so from (3.33) we find the limiting value

$$A_o^\omega = C^\omega + \frac{D^2}{4}(\omega - \sin\omega). \tag{3.35}$$

From (2.6) in Chapter 2 we recall that

$$C^\omega = \frac{D^2}{8}(\omega - \sin\omega),$$

hence (3.35) becomes

$$A_o^\omega = \frac{3D^2}{8}(\omega - \sin\omega) = 3C^\omega. \tag{3.36}$$

This gives a new proof of Theorem 2.1, and (3.36) reduces to (3.1) when $\omega = 2\pi$.

For arclength, the limiting value of (3.34) gives us

$$L_o^\omega = 4D\sin^2\frac{\omega}{4} = 2D(1 - \cos\frac{\omega}{2}), \tag{3.37}$$

a result that will be obtained in a different manner in Chapter 11, (11.16).

Incomplete epicyclogons and hypocyclogons.

Corresponding formulas can be obtained for areas A^p and arclengths L^p of incomplete epicyclogons and hypocyclogons. When a regular n-gon rolls along a regular m-gon, with the edges of the two polygons having equal length, we obtain the following result for area:

Theorem 3.8. *The area of an epicyclogonal or hypocylogonal sector is related to the area of the corresponding cyclogonal sector by*

$$A^p - P_n^p = (1 \pm \frac{n}{m})(A_o^p - P_n^p). \tag{3.38}$$

When we add $A^p - P_n^p$ for the epicyclogonal and hypocyclogonal sectors, the plus and minus terms cancel in (3.38) and we find:

Corollary 3.2. *The sum of differences $A^p - P_n^p$ for the epicyclogonal and hypocyclogonal sectors is $2(A_o^p - P_n^p)$, twice the corresponding result for a cyclogonal sector.*

For arclength we have the following companion results:

Theorem 3.9. *The arclength of an incomplete epicyclogon or hypocyclogon is related to that of the corresponding cyclogon as follows:*

$$L^p = (1 \pm \frac{n}{m})L_o^p. \tag{3.39}$$

Corollary 3.3. *The sum of the epicyclogonal and hypocyclogonal arclengths is $2L_o^p$.*

Corollaries 3.2 and 3.3 will be extended in Section 3.14. When n, m tend to ∞ in such a way that $n/m \to r/R$, Theorem 3.8 gives

$$A^\omega - C^\omega = (1 \pm \frac{r}{R})(A_o^\omega - C^\omega), \tag{3.40}$$

which implies Theorem 2.2. For arclength, Theorem 3.9 yields the corresponding result

$$L^\omega = (1 \pm \frac{r}{R})L_o^\omega, \tag{3.41}$$

which will be obtained differently in Chapter 11, (11.17).

3.10 ARCLENGTH AND AREA OF INVOLUTOGONS

When a taut inelastic string is unwrapped from a point on a given curve, its free end traces a curve called the *involute* of the given curve. Involutes are treated in Chapter 11, but here we introduce the concept of *involutogon*, a curve traced by a point on a straight line that rolls around a fixed polygon. This can be regarded as the limiting case of an epicyclogon when the number of sides of the rolling polygonal disk tends to infinity.

This section treats the simplest case, in which the polygon is a regular m-gon, illustrated in Figure 3.21 with $m = 8$. Here, the moving line (the limiting case of a rolling polygon) is initially along the upper edge of the 8-gon, and the tracing point is initially at a vertex. This can be realized physically by a taut inelastic string unwrapped from the 8-gon. The free end of the string traces an involutogon. Figure 3.21 shows a portion of the involutogon composed of four circular arcs.

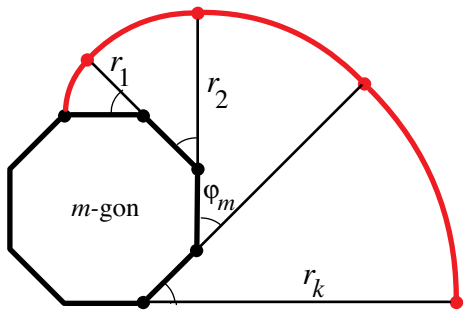


Figure 3.21: An involutogon traced by a point rolling around a regular 8-gon.

We allow the string to turn through p edges of a general regular m-gon and obtain formulas for the arclength L^p traced by the vertex, and for the area A^p of the region swept by the string. If each edge of the m-gon has length a, the involutogon consists of circular arcs of radii $a, 2a, 3a, ..., pa$, each arc subtending the same angle φ_m at each vertex. The length of the kth arc is $ka\varphi_m$ and the area of the corresponding circular sector is $(ka)^2\varphi_m/2$. Consequently, we have

$$L^p = a\varphi_m \sum_{k=1}^{p} k = a\varphi_m \frac{p(p+1)}{2}, \tag{3.42}$$

and

$$A^p = \frac{1}{2}a^2\varphi_m \sum_{k=1}^{p} k^2 = \frac{1}{2}a^2\varphi_m\left(\frac{p^3}{3} + \frac{p^2}{2} + \frac{p}{6}\right). \tag{3.43}$$

We can also express these formulas in terms of the radius R of the circle that circumscribes the m-gon. We have $a = 2R\sin(\pi/m)$ and $\varphi_m = 2\pi/m$, so (3.42) and (3.43) become

$$L^p = p(p+1)\left(\frac{\pi}{m}\sin\frac{\pi}{m}\right)(2R), \tag{3.44}$$

and

$$A^p = (\frac{p^3}{3} + \frac{p^2}{2} + \frac{p}{6})(\frac{\pi}{m}\sin^2\frac{\pi}{m})(4R^2). \tag{3.45}$$

In the limiting case, when $m \to \infty$, the involutogon becomes the involute of a circle of radius R. Let $\alpha = 2\pi p/m$, the central angular portion of the m-gon unwrapped by the string. If the ratio p/m is kept constant as m and p become infinite, then from (3.44) and (3.45) we find that $L^p \to L(\alpha)$ and $A^p \to A(\alpha)$, where

$$L(\alpha) = \frac{1}{2}R\alpha^2, \tag{3.46}$$

and

$$A(\alpha) = \frac{1}{6}R^2\alpha^3. \tag{3.47}$$

These results will be obtained in another way in Chapter 11.

3.11 AREA AND ARCLENGTH OF AUTOGONS

We return to an arbitrary n-gon (not necessarily regular) that rolls around a mirror image of itself, so that in one revolution each edge is made to coincide with a congruent edge. A point z rigidly attached to the rolling n-gon traces a curve we call an *autogon*. Figure 3.20 shows an autogon traced by the vertex of a regular hexagon. Figure 3.22 shows an autogon traced by the point z inside the nonregular quadrilateral that appears in Figure 3.18. (This is an example of a curtate epitrochogon.)

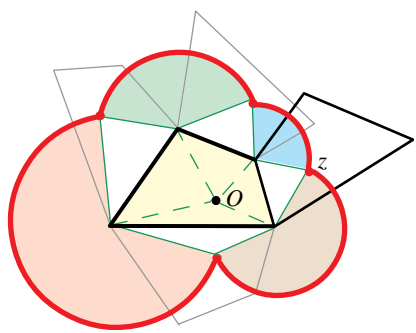


Figure 3.22: An autogon generated by the quadrilateral in Figure 3.18 rolling around a fixed copy of itself. The point O is the mirror image of the tracing point z.

For a general autogon the turning angle at each vertex is twice that of the exterior angle ϕ_k that appears in rolling along a line. Therefore the general formulas for area and arclength given in (3.17) and (3.18) become

$$A_a = P + \sum_{k=1}^{n} \phi_k r_k^2, \tag{3.48}$$

and

$$L_a = 2\sum_{k=1}^{n} \phi_k r_k, \tag{3.49}$$

where the subscripts on the symbols A and L indicate that we are considering autogons. Although there is no general formula for simplifying the sums in (3.48) and (3.49), we can compare the area and arclength of an autogon with the corresponding cyclogonal area A and arclength L as given in (3.17) and (3.18). Comparing (3.48) with (3.17) we see that $A_a - P = 2(A - P)$, so $A_a + P = 2A$. The geometric meaning of this is illustrated in Figure 3.22. The autogon consists of one closed arch surrounding the fixed polygon, so the entire area enclosed inside the autogon, $A+P$, is twice the area of the corresponding cyclogon. Comparing (3.49) with (3.18) we see that $L_a = 2L$, so the arclength of an autogon is twice that of the corresponding cyclogon. Thus we have discovered the following comparison theorems.

Theorem 3.10. *The area of the region inside an autogon is twice that of the corresponding cyclogonal arch obtained by rolling the polygon along a line.*

Theorem 3.11. *The arclength of an autogon is twice that of the corresponding cyclogonal arch obtained by rolling the polygon along a line.*

Both theorems are valid for any location of the tracing point z. They can be visualized geometrically by comparing the autogon in Figure 3.22 with the trochogon in Figure 3.18. Both are traced by the same point z inside the same rolling quadrilateral. Each circular sector in Figure 3.22 subtends twice the angle of the corresponding sector in Figure 3.18 but has the same radius, hence twice the area and twice the arclength. The polygonal region surrounded by the sectors in Figure 3.22 contains the quadrilateral plus the four triangles that provide a dissection of the quadrilateral, so this polygonal region has area twice that of the quadrilateral.

Incomplete autogons.

For incomplete autogons, we simply alter (3.48) and (3.49) by replacing the upper limit n of summation by p. The argument used to prove Theorems 3.10 and 3.11 gives the same type of results for incomplete autogons. Now an incomplete autogonal sector includes the corresponding sector of the fixed polygon. All the sectors have a common vertex at the mirror image O of the tracing point z, as illustrated in Figure 3.22.

Theorem 3.12. *The area of an autogonal sector is twice that of the corresponding cyclogonal sector obtained by rolling the polygon along a line.*

Theorem 3.13. *The arclength of an incomplete autogon is twice that of the corresponding cyclogonal arc obtained by rolling the polygon along a line.*

As limiting cases of these theorems we obtain the following corollaries for an autotrochoid, a trochoid obtained by rolling a smooth curve along a fixed mirror image of itself.

Corollary 3.4. *The area of an autotrochoidal sector of a smooth curve is twice that of the corresponding sector obtained by rolling the curve along a line.*

Corollary 3.5. *The length of an incomplete autotrochoidal arc of a smooth curve is twice that of the corresponding arc obtained by rolling the curve along a line.*

For example, a point rigidly attached to a circular disk traces a autotrochoid (curtate, prolate, or a cardioid) when the disk rolls around an equal disk, and a cycloidal curve (curtate, prolate, or ordinary cycloid) if it rolls along a line. Corollary 3.4 tells us that the area of the autotrochoidal sector is twice that of the corresponding cycloidal sector. Corollary 3.5 tells us that the length of an autotrochoidal arc is twice that of the corresponding cycloidal arc. In particular, the results are valid for full arches. Applications of the corollaries appear in the next two sections.

3.12 ELLIPTIC, HYPERBOLIC, AND PARABOLIC CATENARIES

Elliptic catenary.

Figure 3.23a shows an ellipse rolling along a line. The cycloidal curve traced by one focus F_1 is called an *elliptic catenary*. An elliptic catenary consists of periodically repeating arches, each corresponding to one complete revolution of the ellipse. We wish to determine the arclength s and the area of the ordinate set shown in Figure 3.24a. To do this we roll the ellipse along a fixed copy of itself, through the same length of elliptical arc, as in Figure 3.23b, and apply Corollaries 3.4 and 3.5. In this case the focus traces an autotrochoid that is a circle whose radius R is the length of the major axis of the ellipse.

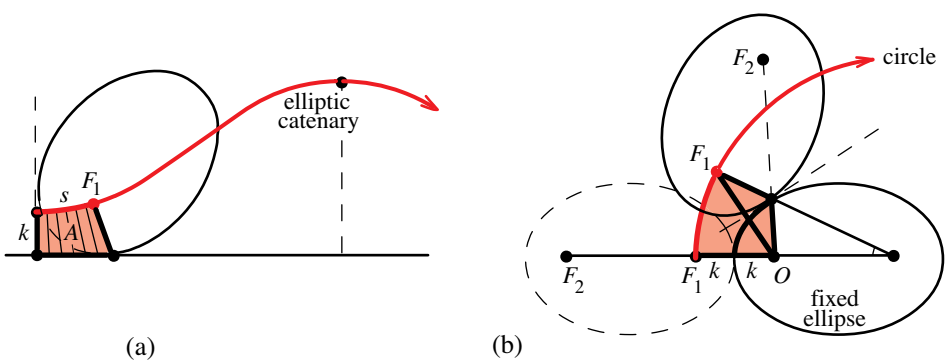


Figure 3.23: (a) When an ellipse rolls along a line, the focus F_1 traces an elliptic catenary. (b) When the ellipse rolls around a fixed copy of itself, the focus F_1 traces a circular arc whose radius R is the length of the major axis.

According to Corollary 3.4, the area of the region bounded by one full arch of the elliptic catenary is half that of the circle of radius R in Figure 3.23b, or $\pi R^2/2$. And by Corollary 3.5, the arclength of the elliptic catenary is half the circumference of the circle, or πR. More generally, the corollaries tell us that when the ellipse in

Figure 3.23b rolls part way around the fixed ellipse, and when the ellipse rolls along a line through the same length of elliptical arc, then the corresponding area A (shaded in Figure 3.23a) is half that of the shaded portion of the circular sector in Figure 3.23b. And the arclength s of the elliptic catenary in Figure 3.23a is half that of the circular arc in Figure 3.23b. The area of the shaded portion of the circular sector in Figure 3.23b is easily calculated. It is the area of the circular sector minus the area of a triangle with edges r, $R - r$, $R - 2k$.

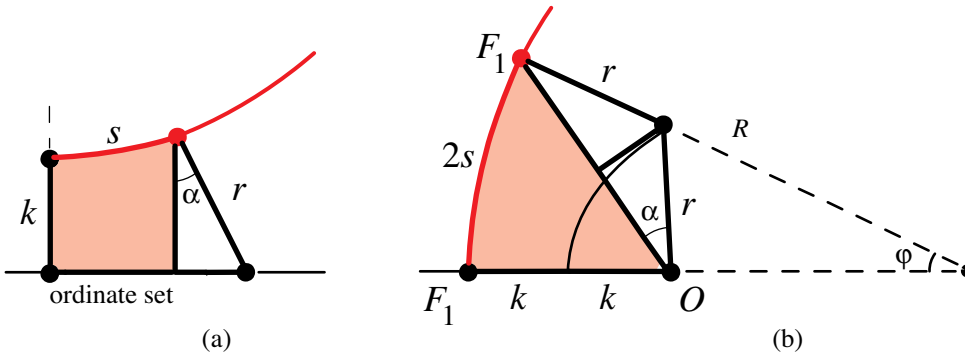


Figure 3.24: Close-up view of Figure 3.23. The area of the ordinate set under the elliptic catenary in (a) is half that of the shaded region in (b).

Figure 3.24, a close-up view of Figure 3.23, shows how to find the area of the ordinate set under the elliptic catenary. The right triangles with hypotenuse r and adjacent angle α are congruent. Consequently, the area of the ordinate set is half the area of the shaded region in Figure 3.24b.

Hyperbolic catenary.

The same problems can be solved for a hyperbola. Figure 3.25a shows one branch of a hyperbola rolling along a line. The focus F_1 enclosed by that branch traces a curve called a *hyperbolic catenary*.

Unlike the ellipse, which can make a complete revolution, the hyperbola can roll through an angle no greater than γ, half the angle between its asymptotes, which is related to the eccentricity e of the hyperbola by $\cos\gamma = 1/e$ (see Figure 7.9). Therefore a hyperbolic catenary is a symmetric arc of limited extent.

Figure 3.25b shows the same hyperbola rolling along a fixed copy of itself, with focus F_1 tracing a circular arc whose radius R is the length of the major axis of the hyperbola.

The same argument used for the elliptic catenary shows that when the hyperbola in Figure 3.22b rolls part way around the fixed hyperbola, and when the hyperbola rolls along a line through the same length of hyperbolic arc, then the corresponding area (shaded in Figure 3.25a) is half that of the shaded region outside of the circular sector in Figure 3.25b. And the arclength s of the hyperbolic catenary in Figure 3.25a is half that of the circular arc in Figure 3.25b. The area of the shaded region

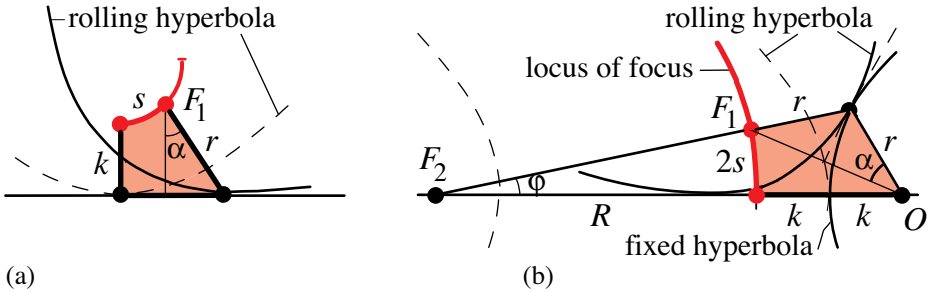


Figure 3.25: (a) When a hyperbola rolls along a line, the focus F_1 traces a hyperbolic catenary. (b) When the hyperbola rolls along a fixed copy of itself, the focus F_1 traces a circular arc whose radius R is the length of the major axis.

in Figure 3.25b is the area of the triangle with edges $r, R+r, R+2k$ minus the area of the circular sector of radius R.

Parabolic catenary.

We also consider the same problems for a parabola. Figure 3.26a shows a parabola rolling around a line. Its focus F traces a symmetric curve of infinite extent called a *parabolic catenary*. When the same parabola rolls along a fixed copy of itself, its focus traces the directrix of the fixed parabola. When the parabola rolls along the line through the same length as the parabolic arc, then the corresponding area A (shaded in Figure 3.26a) is half that of the trapezoid shown in Figure 3.26b.

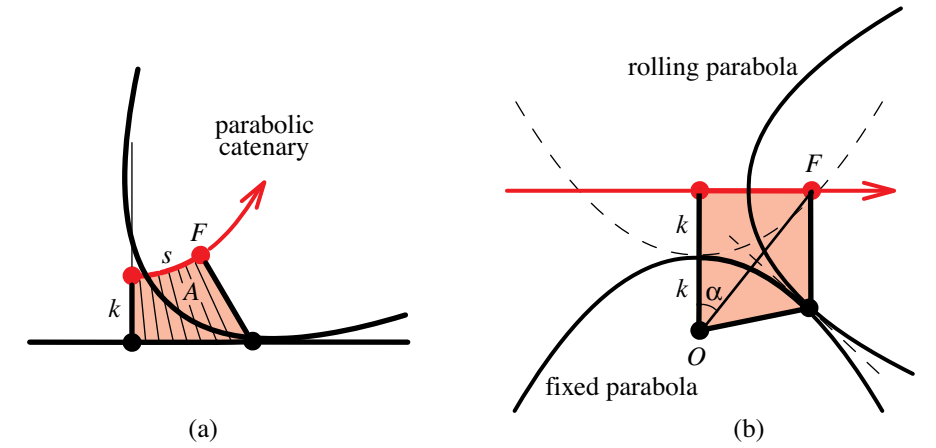


Figure 3.26: (a) A parabola rolling around a fixed copy of itself. The focus F traces the directrix of the fixed parabola. (b) When the same parabola rolls along a line, the focus F traces a parabolic catenary.

Explicit formulas.

Figure 3.27, a close-up view of Figure 3.26, shows how to find the area of the ordinate set under the parabolic catenary. The right triangles with hypotenuse r and adjacent angle α are congruent. Consequently, the area of the ordinate set is half the area of the shaded right triangle in Figure 3.27b. That right triangle has area $2ks$ where s is the arclength of the parabolic catenary, so the area of the

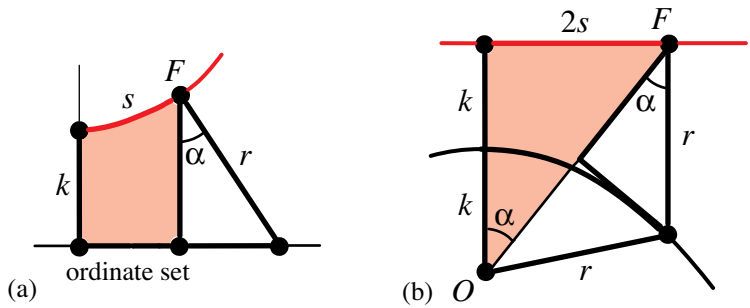


Figure 3.27: Close-up view of Figure 3.26. The area of the ordinate set under the parabolic catenary in (a) is half that of the trapezoid in (b).

ordinate set of the parabolic catenary is given by

$$\text{Area of ordinate set} = ks. \tag{3.50}$$

From the right triangle in Figure 3.24b se see that

$$s = k \tan \alpha. \tag{3.51}$$

As indicated in the diagrams, k denotes the distance from the focus to the vertex of the conic.

In Chapter 11, (11.24), we find that (3.51) gives the natural equation of an ordinary catenary, which is well known as the shape of a uniform flexible chain that hangs under its own weight. Thus, the parabolic catenary is the same as the ordinary catenary. As a consequence of our results concerning autotrochoids we have calculated, in a completely elementary way, both the arclength s and the area of the ordinate set of a catenary. The same results will be obtained by a different method in Chapter 11.

Corresponding results for the arclength and area of the ordinate set of an elliptic catenary in Figure 3.24a can also be obtained in an elementary way as shown above. The arclength $s = R\varphi/2$, half that of the circular arc shown in Figure 3.24b. The angle φ can be determined by applying the law of sines to the triangle with two edges of lengths r and $R - 2k$. The angle opposite that of edge length $R - 2k$ is 2α, hence $(R - 2k) \sin \varphi = r \sin 2\alpha$, which gives

$$s = \frac{1}{2} R\varphi = \frac{1}{2} R \arcsin\left(\frac{r \sin 2\alpha}{R - 2k}\right). \tag{3.52}$$

As $R \to \infty$, $\varphi \to 0$, the ellipse becomes a parabola, and it is easily shown that the limiting value of s in (3.52) is $k \tan \alpha$, in agreement with (3.51).

As already noted, the area of the ordinate set in Figure 3.24a is half the area (that we call S) of the shaded region in Figure 3.24b. The area S, in turn, is the area $R^2\varphi/2$ of the circular sector subtended by φ, minus the area (that we call T) of the triangle with base angle φ and adjacent edges R and $R - 2k$. The altitude of the triangle is $r \sin 2\alpha$, measured from O to the side of length R, so $T = \frac{1}{2} rR \sin 2\alpha$. Hence the area of the ordinate set in Figure 3.24a is given by $\frac{1}{2}(S - T)$. This can be written as

$$\text{Area of ordinate set} = ks + \frac{1}{4} R(R - 2k)(\varphi - \sin \varphi).$$

An alternative form is

$$\text{Area of ordinate set} = \frac{1}{2} kR \sin \varphi + \frac{1}{4} R^2(\varphi - \sin \varphi). \tag{3.53}$$

It is easily shown that as $R \to \infty$ and $\varphi \to 0$ in such a way that $R\varphi/2 = s$, the ellipse becomes a parabola, and the limiting value of (3.53) is ks, in agreement with (11.25), which gives the area of the ordinate set of an ordinary catenary.

For a hyperbolic catenary, it is easily shown that the formulas corresponding to (3.52) and (3.53) are, respectively,

$$s = \frac{1}{2} R\varphi = \frac{1}{2} R \arcsin\left(\frac{r \sin 2\alpha}{R + 2k}\right),$$

$$\text{Area of ordinate set} = \frac{1}{2} kR \sin \varphi - \frac{1}{4} R^2(\varphi - \sin \varphi).$$

To the best of our knowledge, the area of the ordinate set of an elliptic or hyperbolic catenary has not been previously treated, except for the full area in the elliptic case. It should be noted that these elementary formulas are expressed in terms of the angle of rotation.

3.13 PEDAL CURVES AND STEINER'S THEOREMS

Some roulettes are pedal curves, which are defined as follows. Given a smooth plane curve Γ and a point z not on Γ called a pedal point, let p denote the foot of the perpendicular from z to a typical tangent line to Γ. The locus of all such points p constructed for all tangent lines to Γ is called the pedal curve of Γ with respect to z. In Section 1.15 we introduced the limaçon of Pascal as the pedal curve of a circle. In particular, if the pedal point is on the circle, the limaçon is a cardioid.

The 19th century geometer Jakob Steiner discovered a remarkable property relating the area of a roulette with that of its pedal curve, and another relating their arclengths.

Theorem 3.14. (Steiner's first theorem) *When a smooth closed curve Γ rolls along a straight line, the area of the region between one full arch of the roulette traced by a point z attached to Γ and the straight line is twice the area of the region enclosed by the pedal curve of Γ with z as pedal point.*

Theorem 3.15. (Steiner's second theorem) *The length of an arc of the roulette is equal to the arclength of the pedal curve.*

We will deduce Steiner's theorems from Corollaries 3.4 and 3.5. But first we obtain comparison results relating areas and arclengths of autogons with those of their pedal curves.

Autogons and pedal curves.

Because a polygon has only a finite number of tangent lines (one for each edge) the usual definition of pedal curve would produce only a finite set of points as the pedal curve. Hence we need to extend the concept of pedal curve so that it applies to polygons. The problem is to introduce a suitable replacement for a tangent line at a vertex of a polygon. We do this as follows.

Each vertex v of a polygon is the intersection of two consecutive edges. From a given pedal point z inside or outside the polygon, perpendiculars to the edges intersect the polygon at points p and q, say. In Figure 3.28a, z is inside.

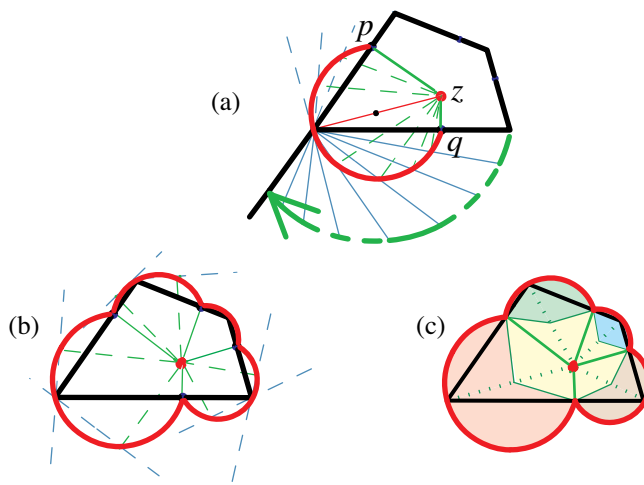


Figure 3.28: The pedal curve of a quadrilateral with a pedal point z inside.

Imagine rotating a line through one edge about this vertex through the exterior angle at that vertex until it reaches the line of an adjacent edge. Each intermediate position of the rotating line can play the role of a tangent line, and a perpendicular can be drawn from the pedal point z to each such line. The locus of the feet of the perpendiculars will lie on a circle (passing through the common vertex) whose diameter d is the distance from z to that vertex. The circular arc joining p and q is, by definition, the portion of the pedal curve contributed by that vertex. So the pedal curve of an n-gon consists of n circular arcs, one for each vertex. The diameter of each arc is the distance from the pedal point to the corresponding vertex. The pedal curve of the quadrilateral in Figure 3.28b consists of four circular arcs. For a convex polygon, no point of the pedal curve lies inside the polygon.

Now take an arbitrary n-gon and roll it around its fixed mirror image so that in one revolution each edge is made to coincide with a congruent edge. A point z rigidly attached to the rolling n-gon traces an autogon. The pedal curve to this autogon with pedal point z will be a scaled copy of the autogon with a scaling factor of $1/2$. This fact is known for smooth curves but it holds for arbitrary n-gons as well. The reason is that each point on the pedal curve is the foot of a perpendicular from the pedal point z and therefore must be at the midpoint of the extended perpendicular that joins z to its mirror image. Consequently, we have the following comparison results, not only for complete autogons and their pedal curves, but for incomplete autogonal sectors as well. They can be regarded as companion results to Corollaries 3.4 and 3.5:

Theorem 3.16. *The area of an autogonal sector traced by z is four times that of the corresponding sector of the pedal curve having pedal point z.*

Theorem 3.17. *The arclength of an autogonal arc traced by z is twice that of the corresponding portion of the pedal curve having pedal point z.*

These can be visualized geometrically by comparing the shaded regions inside the pedal curve in Figure 3.28c with the autogon in Figure 3.22. The same point z traces the autogon in Figure 3.22 and is the pedal point in Figure 3.28. Each circular sector swept out by z in Figure 3.22 subtends the same angle as in Figure 3.28b but has twice the radius. So the area of the sector in Figure 3.22 is four times that in Figure 3.28c, and its arclength is twice that in Figure 3.28c. And the polygonal region surrounded by the circular sectors in Figure 3.22 has linear dimensions twice those in Figure 3.28c, so its area is four times as large. In particular, we have the following examples:

1. *The area of a sector inside a limaçon of Pascal* (*or a cardioid*) *is twice that of the corresponding curtate or prolate cycloidal sector* (*or cycloidal sector*) *obtained by rolling the same disk along a line.*

2. *The length of an arc of the limaçon of Pascal* (*or a cardioid*) *is twice that of the corresponding curtate or prolate cycloidal arc* (*or cycloid*) *obtained by rolling the same disk along a line.*

These results for the area of a full limaçon of Pascal and for the length of an arbitrary arc were known to Steiner.

New proofs of Steiner's theorems.

Now we use Theorems 3.16 and 3.17 together with Corollaries 3.4 and 3.5 to deduce Steiner's theorems for smooth curves. By Corollary 3.4, the area of a cyclogonal arch traced by a point z attached to an n-gon rolling along a straight line is half that of the corresponding autogon, hence, by Theorem 3.16, it is twice that of the pedal curve of the n-gon with pedal point z. This implies Steiner's first theorem for any smooth curve that is the limit of n-gons. Steiner's second theorem follows similarly from Corollary 3.5 and Theorem 3.17.

3.14 REDUCTION FORMULAS FOR ARCLENGTHS AND AREAS

This section exploits the fact that some arclength and area calculations for trochogonal curves can be reduced to those of cyclogons generated by the same rolling polygon.

Curves traced by a point attached to a rolling regular polygon.

When a regular n-gon rolls along a line through p of its edges, where $p \leq n$, a vertex of the n-gon traces an incomplete cyclogon of arclength L_o^p. When the same n-gon rolls through p of its edges along a regular m-gon, with the edges of the two polygons having equal lengths, the same vertex traces an incomplete epicyclogon or hypocyclogon of arclength L^p. Theorem 3.9 reduces the arclength L^p to the cyclogonal arclength L_o^p by the simple relation

$$L^p = (1 \pm \frac{n}{m})L_o^p, \tag{3.54}$$

with the + sign for the epicyclogon and the − sign for the hypocyclogon.

The same type of reduction holds for prolate or curtate epicyclogons or hypocylogons traced by a point z attached to the rolling n-gon. This is easily seen by replacing (3.28) by the more general formula

$$L^p(z) = (\frac{2\pi}{n} \pm \frac{2\pi}{m}) \sum_{k=1}^{p} r_k = (1 \pm \frac{n}{m})(\frac{2\pi}{n} \sum_{k=1}^{p} r_k), \tag{3.55}$$

where we have written $L^p(z)$ to indicate dependence on z (because now r_k depends on z). The last factor in (3.55) represents the length $L_o^p(z)$ of the cyclogonal curve traced by the same point z when the n-gon rolls along a line. Thus reduction formula (3.55) becomes

$$L^p(z) = (1 \pm \frac{n}{m})L_o^p(z), \tag{3.56}$$

which has the same form as the special case in (3.54) when z is a vertex. In general, there is no direct way to determine either arclength $L^p(z)$ or $L_o^p(z)$, but the reduction formula establishes a simple relation between them.

The same analysis shows that the corresponding reduction formula for areas in Theorem 3.8 can be generalized when the tracing point is an arbitrary point z attached to the rolling n-gon. In this case we have an extension of (3.38):

$$A^p(z) - P_n^p(z) = (1 \pm \frac{n}{m})(A_o^p(z) - P_n^p(z)), \tag{3.57}$$

where $P_n^p(z)$ is the area of the accumulated triangular footprints of the rolling n-gon like those in Figure 3.18.

Sum of arclengths and areas of complementary trochogonal curves.

For an epicycloid obtained by a disk of radius r rolling outside a fixed circle of radius R, there is a corresponding hypocycloid obtained by a disk of the same radius r rolling inside. In Chapter 2 we called such curves *complementary* and observed that the sum of the areas of complementary arches is six times the area of the rolling disk, which is twice the area of the cycloidal arch traced by the same disk rolling on a line. (Recall Corollary 2.3 and Figure 2.13.) This fact is remarkable because it does not depend on the radius R of the fixed circle. Theorem 2.7 extended this to complementary epitrochoidal and hypotrochoidal curves obtained by rolling a disk on opposite sides of a general base curve Γ. The sum of the areas of an epitrochoidal sector and its complementary hypotrochoidal sector does not depend on Γ and is twice the area of the corresponding cycloidal sector. In Chapter 11, Section 11.5, it is shown that a similar relation holds for arclengths, their sum being twice that of the corresponding cycloidal arclength. Therefore it is not surprising that similar relations hold for complementary trochogonal curves traced by not only the vertex of a regular polygon but also by a point z attached to a general convex rolling polygon, not necessarily regular.

The reason for this is easily explained by comparing Figures 3.29a and b. Figure 3.29a shows a polygon and its mirror image rolling on opposite sides of a straight line Γ_0. A vertex V and its mirror image V' trace complementary circular arcs with equal central angles ϕ at the pivot vertex.

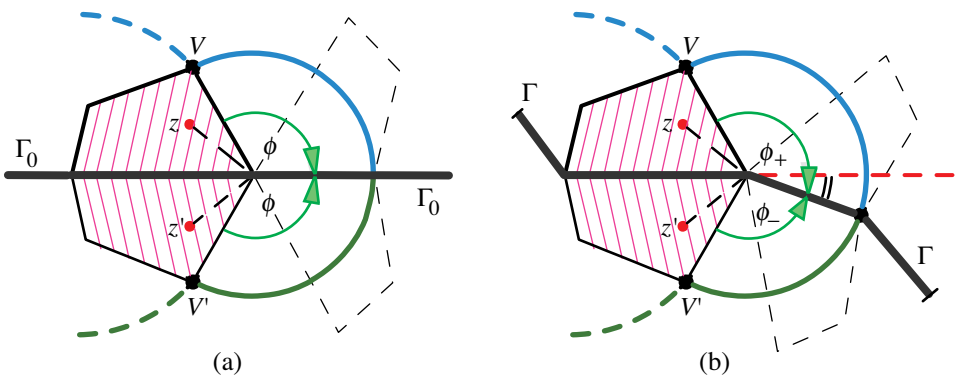


Figure 3.29: The two central angles in (a) have the same sum as those in (b).

In Figure 3.29b, Γ_0 is replaced by a polygonal track Γ which is a broken version of the rolling polygon's boundary, configured so that the polygon and its mirror image can roll along opposite sides of Γ. The angles between edges of Γ are arbitrary. Now vertex V and its mirror image V' trace complementary circular arcs with central angles ϕ_+ and ϕ_- at the pivot vertex. They are not equal if there is a bend in the track, but at each pivot vertex their sum is given by $\phi_+ + \phi_- = 2\phi$, where ϕ is the corresponding turning angle in Figure 3.29a.

Consequently, the sum of the arclengths of the complementary circular arcs accumulated as the polygon rolls along any number of edges does not depend on

how Γ bends from vertex to vertex, and it is twice the arclength accumulated when the polygon rolls on Γ_0 along the same number of edges. And the same is true for the sum of areas of the complementary trochogonal arches accumulated during an incomplete rotation, whether or not the triangular footprints of the rolling polygon are included.

A similar analysis reveals that the same properties hold for arclengths and areas of complementary trochogons traced by a point z and its mirror image z'.

The foregoing claims are summarized in the following two theorems, in which the tracing point z is attached to an arbitrary convex polygon that rolls along a polygonal track Γ as described above. Point z and its mirror image trace complementary trochogonal arcs.

Theorem 3.18. *The sum of lengths of the complementary arcs is independent of* Γ *and is twice the length of the arc obtained when the polygon rolls along a line.*

Theorem 3.19. *The sum of areas of the complementary trochogonal sectors* (*with or without footprints*) *is independent of* Γ *and is twice the area of the corresponding sector obtained when the polygon rolls along a line.*

As limiting cases of these theorems we obtain the following corollaries for complementary trochoids obtained when a smooth curve and its mirror image roll on opposite sides of a curve Γ.

Corollary 3.6. *The sum of arclengths of complementary trochoids is independent of* Γ *and is twice the arclength of the trochoid obtained when* Γ *is a straight line.*

Corollary 3.7. *The sum of areas of complementary trochoidal sectors is independent of* Γ *and is twice the area of the corresponding sector obtained when* Γ *is a straight line.*

The results for autogons in Theorems 3.12 and 3.13 and for autotrochoids in Corollaries 3.4 and 3.5 can be regarded as special cases.

NOTES ON CHAPTER 3

This chapter is adapted from material published in [5], [9], and [12]. Sections 3.9 and 3.14, which deal with areas and arclengths of incomplete trochogonal arches, have not been previously published. The same is true of Section 3.12 on elliptic, hyperbolic, and parabolic catenaries, and of Section 3.10, which deals with involutogons. The limiting cases in Section 3.9, when the polygonal disks become circular, are treated differently in Chapter 2 for areas, and in Chapter 11 for arclength. Also, the results of Section 3.10 for the limiting case of an involutogon are obtained differently in Chapter 11.

Animation showing the generation of elliptic, hyperbolic, and parabolic catenaries, together with mechanisms for tracing these curves, can be seen on the web site

www.mathcurve.com/courbes2d/delaunay/delaunay.shtml

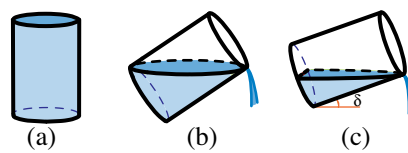
Chapter 4

CIRCUMGONS AND CIRCUMSOLIDS

These problems can be easily solved by the methods developed in this chapter. The reader may wish to try solving them before reading the chapter.

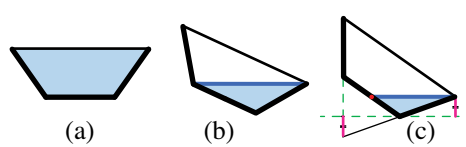
When a cylindrical glass full of water (a) is tipped as in (b), exactly half the volume of water remains in the glass. In (c), the glass is tipped further so that the surface of the water bisects the circular base.

Calculate the fraction of the volume of water that remains in the glass if angle $\delta = 30°$ *in* (c).



Now the glass has the shape of the frustum of a cone as shown below. It is initially full of water in (a), and is tipped in (b) so that the horizontal level of the water first touches the circular base. In (c) it is tipped further so that the water level bisects the circular base. The shape of the glass is determined by (c), where one slanted edge is vertical, and the two marked vertical segments have equal lengths.

What fraction of the full volume remains in the glass in configuration (b) *and in configuration* (c)?



CONTENTS

4

Every triangle circumscribes a circle called the incircle, whose radius is called the inradius and whose center is called the incenter. A polygon with more than three edges may or may not circumscribe a circle. Those that do are examples of what we call circumgons. Each has an inradius and an incenter. Circumgons include all triangles, all regular polygons, some special irregular polygons and even nonconvex polygons (such as star-shaped polygons), and other plane figures composed of line segments and circular arcs. This chapter shows that all circumgons share common properties regarding area-perimeter ratios and centroids. For example, the ratio of the area of a region bounded by a circumgon to its semiperimeter is equal to its inradius (just as the ratio of the area of a circular disk to its semiperimeter is its radius). Also, the area centroid of a region bounded by a circumgon and the centroid of its boundary curve are collinear with the incenter, at distances in the ratio 2:3 from the incenter, as in the case of a triangle. Corresponding results are derived for circumgonal rings, plane regions lying between two similar circumgons. These rings have constant width. The ratio of the area to the semiperimeter of such a ring is equal to the constant width. Relations between the area centroid of a circumgonal ring with the centroid of its boundary are also given.

The results are extended to n-space by introducing circumsolids, solids that circumscribe an n-sphere. Each circumsolid has an incenter and an inradius. For $n = 3$ they include tetrahedra, regular polyhedra, some special irregular polyhedra and even nonconvex polyhedra (such as stellated polyhedra), and many other solids whose faces can be cylindrical, conical, or spherical, as well as planar. All circumsolids in n-space share a common property: the ratio of volume to outer surface area is $1/n$ times the inradius. Profound implications of this property are given in Chapter 13. Also, the volume centroid of a circumsolid and the centroid of its outer boundary surface area are collinear with the incenter, at distances in the ratio $n/(n + 1)$ from the incenter.

Circumsolids offer a rich variety of applications, as shown by examples in 3-space that include star-like circumsolids such as stellated dodecahedra, and intersections of circumsolids. One application of the volume-surface area

ratio shows that a plane through the incenter of a circumsolid divides it into two smaller solids whose surface areas are equal if, and only if, their volumes are equal. Another yields (without integration) the volume of the solid of intersection of a right circular cone and an orthogonal circular cylinder having the same insphere. A limiting case is the classical Archimedean result on intersecting cylinders.

The chapter also treats circumsolid shells in 3-space, solids lying between two similar circumsolids with the same incenter. They have constant thickness, and the ratio of volume to mixed average surface area is one-third of the constant thickness. This implies far-reaching extensions of the classical Egyptian and prismoidal formulas to nonplanar surfaces.

4.1 INTRODUCTION

We begin by generalizing Archimedes' striking discovery concerning the area of a circular disk, which for our purposes we prefer to state as follows:

Theorem 4.1. (Archimedes) *The area of a circular disk is one-half the product of its perimeter and its radius.*

Expressed as a formula, this becomes

$$A = \frac{1}{2}Pr, \tag{4.1}$$

where A is the area, P is the perimeter, and r is the radius of the disk. First we extend (4.1) to a large class of plane figures circumscribing a circle that we call circumgons, defined in Section 4.2. They include arbitrary triangles, all regular polygons, some irregular polygons, and other figures composed of line segments and circular arcs. Examples are shown in Figures 4.1 through 4.4. Section 4.3 treats circumgonal rings, plane regions lying between two similar circumgons. These rings have a constant width that replaces the radius in the extension of (4.1). We show that all rings of constant width are necessarily circumgonal rings. We also describe centroidal relations associated with circumgons and circumgonal rings. A special case is a result of Archimedes on the centroid of a trapezoid.

In Section 4.6 we provide corresponding results for 3-space, in which we generalize another discovery of Archimedes:

Theorem 4.2. (Archimedes) *The volume of a sphere is one-third the product of its surface area and its radius.*

We extend this result to a class of general solids called circumsolids, a three-dimensional extension of circumgons. The faces of a circumsolid can be curved as well as planar. The curved faces can be cylindrical, conical, or spherical. Each circumsolid circumscribes a sphere, and we provide a direct extension of Theorem 4.2 for any circumsolid. Extensions to n-space for $n \geq 3$ are also indicated.

Many new results emerge as applications. For example, we determine the volume of the solid of intersection of a cone and cylinder without using calculus. We also generalize several classical centroid relations, as well as the Egyptian and prismoidal formulas for volume.

4.2 CIRCUMGONS

To pave the way for the general definition of a circumgon, we begin with some examples. The prototype is a triangle; it circumscribes a circle whose center is the point of intersection of the three angle bisectors. By dividing a triangle into three smaller triangles with a common vertex at the center of the inscribed circle, we easily see that (4.1) holds for any triangle of area A and perimeter P, where r is the radius of the inscribed circle.

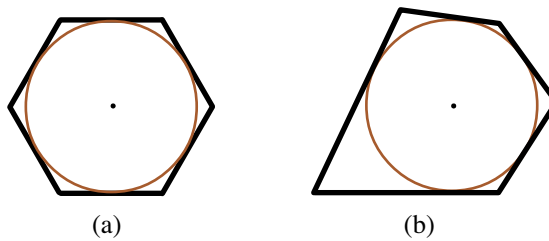


Figure 4.1: Examples of circumgons: (a) a regular hexagon and (b) a nonregular pentagon.

A polygon with more than three edges may or may not circumscribe a circle. We are interested in those that do, because they provide examples of circumgons. Every regular polygon is a circumgon, but there are also nonregular circumgons, as illustrated in Figure 4.1b. Like a triangle, any polygon circumscribing a circle is a circumgon. The inscribed circle is called the *incircle*, its radius is called the *inradius*, and its center is called the *incenter*. Bisectors of the interior angles of a circumgon intersect at the incenter. By dividing the polygon into triangles with one common vertex at the incenter it is easily seen that (4.1) holds for every circumgon whose boundary is a convex polygon.

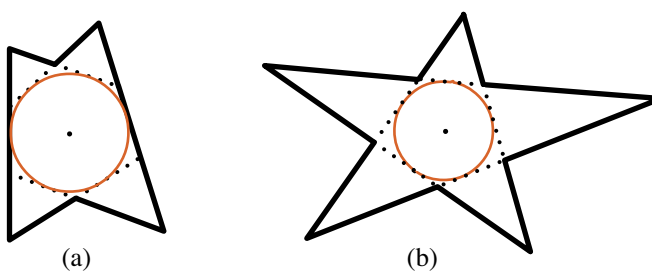


Figure 4.2: The area of each circumgonal region is half its perimeter times its inradius.

We will also extend (4.1) to more general circumgons, not necessarily convex, such as the polygon in Figure 4.2a, or the star-shaped polygon in Figure 4.2b, and to more general polygonal shapes, not necessarily closed, such as the example in Figure 4.4a. It may seem surprising that nonconvex polygons can circumscribe a circle. It's true that our examples are not ordinary garden-variety circumscribing polygons, but

when viewed appropriately, they do circumscribe a circle. For example, in Figure 4.2a only two edges of the polygon are tangent to the incircle. The other four edges do not even touch the incircle, but their extensions, shown by dotted lines, are tangent to the incircle. In Figure 4.2b, none of the edges of the pentagram touches the incircle, but each extended edge, shown by dotted lines, is tangent to the incircle.

Building blocks of a circumgonal region. The definition of a general circumgonal region will be formulated in terms of simpler elements called *building blocks*, defined as follows.

Start with a circle, and consider a triangular region with one vertex at the center and with side opposite the vertex lying on a line tangent to the circle. We call this triangular region a building block of a circumgonal region; the side opposite the center on the tangent line is called the *outer edge* of the block. An example is shown in Figure 4.3a. The circle is the incircle, its radius is the inradius, and its center is the incenter. Because a circular arc can be regarded as a limiting case of circumscribing polygons, we also allow any sector of the incircle to be a building block of a circumgonal region, with its outer edge being the circular arc, as shown in Figure 4.3b. Thus, the area of each building block, whether it is a triangular region or a circular sector, is equal to half the length of its outer edge times the inradius.

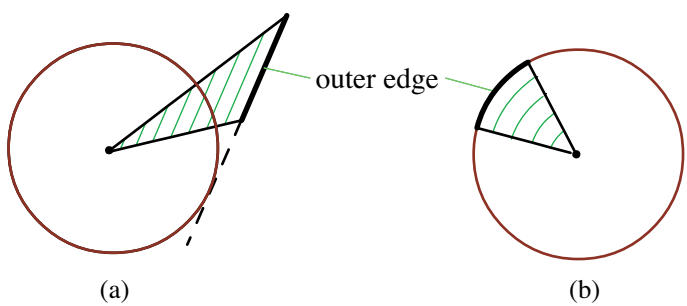


Figure 4.3: A building block of a circumgonal region is either (a) a triangular region or (b) a circular sector. Its perimeter is the length of the outer edge.

To extend Theorem 4.1, we simply define the perimeter of the block to be the length of its outer edge. This gives us:

Theorem 4.3. *The area of a circumgonal building block is equal to half the product of its perimeter and its inradius.*

Definitions of circumgonal region and circumgon. A *circumgonal region* is the union of a finite set of nonoverlapping building blocks having the same incircle. The union of the corresponding outer edges is called a *circumgon*; the sum of the lengths of the outer edges is called the *perimeter* of the circumgon.

The perimeter of a circumgon, as just defined, is not its perimeter in the usual Euclidean sense unless the circumgon is closed.

The definition immediately gives an extension of Theorem 4.3:

Theorem 4.4. *The area of a circumgonal region is equal to half the product of its perimeter and its inradius.*

Both Theorems 4.3 and 4.4 are described by the formula used for Theorem 4.1:

$$A = \frac{1}{2}Pr, \tag{4.2}$$

where A is the area, P is the perimeter, and r is the inradius of the circumgon. Two examples satisfying (4.2) are shown in Figure 4.4.

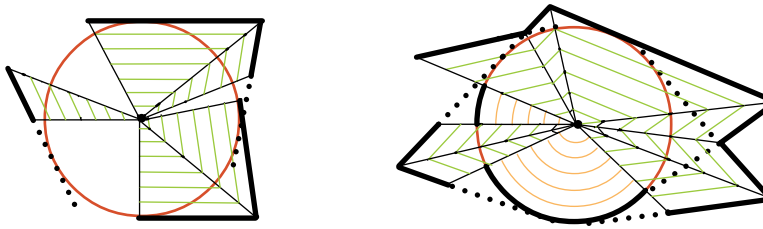


Figure 4.4: More examples of circumgonal regions: the area of each is half its perimeter times its inradius.

4.3 CIRCUMGONAL RINGS

Figure 4.5 (right) shows a simple type of ring, the region between two similar nonoverlapping simple closed curves with similarity ratio λ, where $0 < \lambda < 1$. We call λ the *size factor* because it determines the size of the inner curve relative to the outer one. If the outer region has perimeter P_0 and area A_0, the inner region has perimeter λP_0 and area $\lambda^2 A_0$, regardless of the choice of center of scaling.

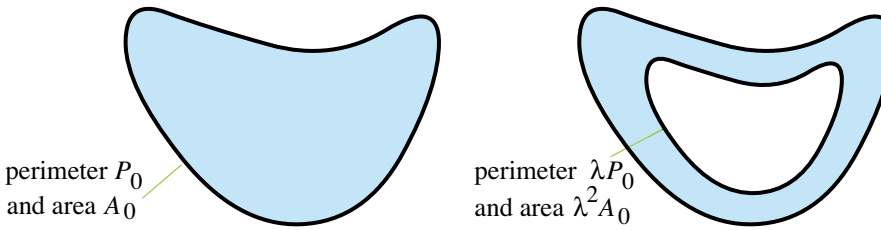


Figure 4.5: A simple closed curve with perimeter P_0 and area A_0 used to form a ring with size factor λ.

We are interested in rings formed by scaling a circumgonal region from its incenter. The inner and outer boundaries need not be closed curves because the circumgons need not be closed.

For a general ring the perpendicular distance between the boundary curves need not be constant, even if portions of the boundaries are parallel, as in the case of two similar rectangles. But if the ring is formed by scaling a circumgonal region from its incenter, it is easy to show that the perpendicular distance between corresponding parallel segments (or circular arcs) is a constant, which we call the *width* of the ring.

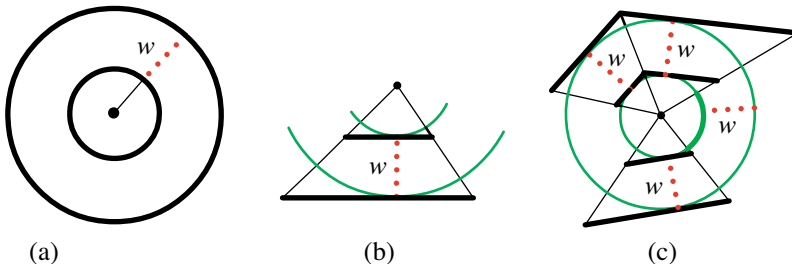


Figure 4.6: Examples of circumgonal rings. The annulus in (a) and the trapezoid in (b) are extreme cases.

Figure 4.6 shows examples of circumgonal rings. The circular annulus in (a) and the trapezoid in (b) are extreme cases. A more general example is shown in (c). In each case, the constant width w is the perpendicular distance between its parallel edges. It is also true that circumgonal rings are the only rings having constant width. In fact, we have:

Theorem 4.5. (a) *A circumgonal ring formed by scaling a circumgonal region from its incenter has constant width.*

(b) *Conversely, consider a ring formed by two similar contours, where the outer contour consists of a finite set of line segments and circular arcs. If the ring has constant width, then it is necessarily a circumgonal ring.*

Proof. The proof of (a) is an easy exercise, which shows that the constant width w is given by

$$w = (1 - \lambda)r,$$

where r is inradius of the larger circumgon and $\lambda < 1$ is the scaling factor.

To prove (b), refer to Figure 4.7, which shows a trapezoidal portion of the ring formed by parallel line segments AB and $A'B'$ having w as the perpendicular distance between them. The intersection O of the lines through AA' and BB' is the center of similarity, and $OA' = \lambda OA$, where λ (< 1) is the scaling factor. Let Q be the foot of a perpendicular from O to the line through AB. The circle with center O and radius OQ is tangent to the line through AB, so AB is an outer edge of a circumgon with incenter O and inradius OQ. By similarity, the point Q' on OQ satisfies $OQ' = \lambda OQ$, and the circle with center O and radius OQ' is tangent to the line through $A'B'$, so $A'B'$ is an outer edge of a circumgon with incenter

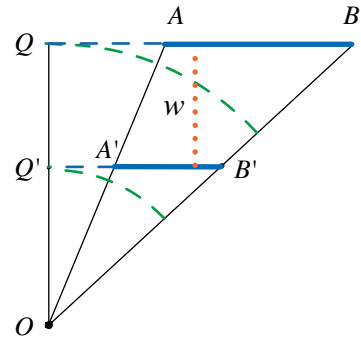


Figure 4.7: Illustrating the proof that every ring of constant width is a circumgonal ring.

O and inradius OQ'. But $w = OQ - OQ' = (1-\lambda)OQ$, hence $OQ = w/(1-\lambda)$ and $OQ' = \lambda w/(1-\lambda)$. Thus the inradii and incenter O are completely determined by the width w and the scaling factor λ, as given above. This means that every trapezoidal portion of the ring circumscribes the same pair of circles. Consequently, the entire polygonal part of the ring is circumgonal with incenter O. The proof is even simpler for each portion of the ring that is a circular sector (of width w).

The next result extends Theorem 4.4 to circumgonal rings.

Theorem 4.6. *The area of a circumgonal ring is equal to half the product of its perimeter and its (constant) width.*

This can also be expressed as a formula resembling (4.2):

$$A = \frac{1}{2}Pw, \tag{4.3}$$

where A is the area of the ring, P is its total perimeter, and w is its constant width.

Proof. If the outer boundary has perimeter P_0 and encloses a region of area A_0, then the ring has area $A = (1-\lambda^2)A_0$ and total perimeter $P = (1+\lambda)P_0$. For a circumgonal ring with inradius r of the larger circumgon we have $A_0 = P_0 r/2$, from which we find that

$$A = (1-\lambda)(1+\lambda)P_0 r/2 = (1-\lambda)Pr/2 = Pw/2,$$

as asserted. It is not surprising that (4.3) gives the formula for the area of a circular ring (Figure 4.6a). But it is reassuring to learn that (4.3) also becomes the well-known formula for the area of a trapezoid, average base times altitude. In fact, in Figure 4.6b half the perimeter of the ring is the average length of the two parallel edges, and the width of the ring is the altitude of the trapezoid.

4.4 CENTROIDS OF CIRCUMGONAL REGIONS

This section derives a simple but surprising relation between the area centroid of a circumgonal region and the centroid of its boundary. Specifically, denote by $\boldsymbol{C}(A)$

the vector from the incenter O to the area centroid, and by $\boldsymbol{C}(B)$ the vector from O to the centroid of the boundary curve (with respect to arclength). Figure 4.8b illustrates these for a triangle. We will prove that, for a circumgon, the location of one of the centroids determines the location of the other. In fact, we have:

Theorem 4.7. *The area centroid $\boldsymbol{C}(A)$ of a circumgonal region and the centroid $\boldsymbol{C}(B)$ of its boundary are collinear with the incenter, and are related by*

$$\boldsymbol{C}(B) = \frac{3}{2}\boldsymbol{C}(A). \tag{4.4}$$

Proof. A classical result of Archimedes states that the area centroid of a triangle is at the intersection of its medians. It is also known that the distance from each vertex to the centroid is two-thirds the length of the median from that vertex. (For proofs, see Chapter 12.) Apply this to the triangular block with incenter O and outer edge of length a shown in Figure 4.8a. In vector notation, $\boldsymbol{C}(A) = (2/3)\boldsymbol{C}(B)$, where $\boldsymbol{C}(B)$ is the midpoint of the outer edge. Hence

$$\boldsymbol{C}(B) = \frac{3}{2}\boldsymbol{C}(A), \tag{4.5}$$

which proves (4.4) for a triangular block.

Now take a polygonal circumgon with triangular building blocks having outer edges of lengths a_1,..., a_n, a common vertex at the incenter O, and respective areas A_1,..., A_n. Figure 4.8b shows the case of a triangle.

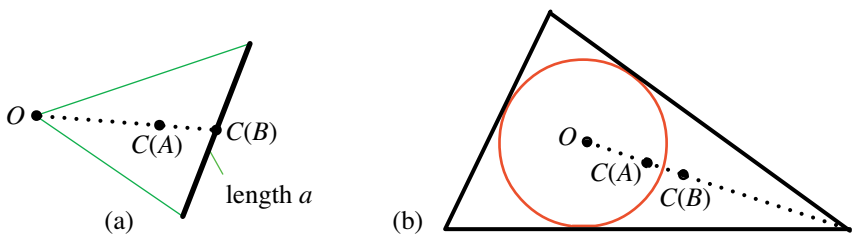


Figure 4.8: Centroid $\boldsymbol{C}(A) = (2/3)\boldsymbol{C}(B)$ for (a) a triangular block, and (b) for any triangle with incenter O.

Denote by $\boldsymbol{C}(A_1)$,..., $\boldsymbol{C}(A_n)$ the corresponding vectors from the incenter O to the area centroid of each triangular block. The area centroid of their union is at the point described by the vector

$$\boldsymbol{C}(A) = \frac{\sum_{k=1}^{n} A_k \boldsymbol{C}(A_k)}{\sum_{k=1}^{n} A_k}. \tag{4.6}$$

In (4.6), write $A_k = a_k r/2$, where r is the inradius. The common factor $r/2$ cancels, and (4.6) becomes

$$\boldsymbol{C}(A) = \frac{\sum_{k=1}^{n} a_k \boldsymbol{C}(A_k)}{\sum_{k=1}^{n} a_k}. \tag{4.7}$$

On the other hand, the vector $\boldsymbol{C}(B)$ from O to the centroid of the boundary is

$$\boldsymbol{C}(B) = \frac{\sum_{k=1}^{n} a_k \boldsymbol{C}(B_k)}{\sum_{k=1}^{n} a_k},$$

where $\boldsymbol{C}(B_k)$ denotes the vector from O to the midpoint of the kth outer edge. Apply (4.5) to each triangular block to find that $\boldsymbol{C}(B_k) = (3/2)\boldsymbol{C}(A_k)$. Use this in the last equation and compare with (4.7) to obtain (4.4) for a polygonal circumgon.

Because a circular arc can be regarded as a limiting case of a circumscribing polygon, (4.4) also holds for circumgons that include circular arcs as part of their boundaries.

We can also deduce (4.4) for a circular sector in a different manner. It is known (see Chapter 12) that the area centroid of a circular sector of radius r subtending a central angle 2α lies on the radial line that bisects the sector at a distance $(2/3)r(\sin\alpha)/\alpha$ from the center, and the centroid of the outer arc is at a distance $r(\sin\alpha)/\alpha$ from the center. Consequently, (4.4) holds for every circular sectorial building block of a circumgon.

In particular, (4.4) holds for any triangle, and also for any polygon circumscribing a circle. Because these two cases are so basic, we restate them here as corollaries:

Corollary 4.1. (a) *The area centroid $\boldsymbol{C}(A)$ of a triangle and the centroid $\boldsymbol{C}(B)$ of its boundary are collinear with the incenter and are related by*

$$\boldsymbol{C}(B) = \frac{3}{2}\boldsymbol{C}(A). \tag{4.8}$$

(b) *The same relation holds for a polygon circumscribing a circle.*

The results for these two classical cases are so simple that we thought they must surely be known. If so, they are well hidden because we could find neither of them in the literature.

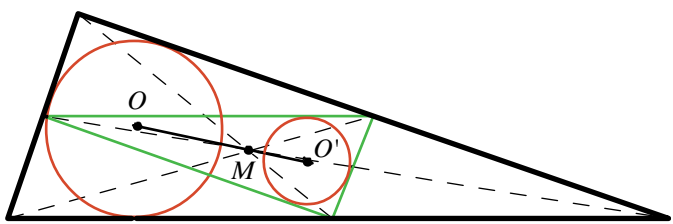


Figure 4.9: Another argument showing that $\boldsymbol{C}(B) = (3/2)\boldsymbol{C}(A)$ for a triangle.

Another derivation of Corollary 4.1(a) can be given by referring to Figure 4.9. It is known that the centroid of the boundary of a triangle is the incenter O' of the medial triangle shown. Both triangles have common median lines, hence a common area centroid at point M. The incenter O of the larger triangle is collinear with O' and M, and its inradius is twice that of the smaller triangle, so $OM = 2MO'$. Consequently $OO' = OM + MO' = (3/2)OM$, in agreement with (4.8).

4.5 CENTROIDS OF CIRCUMGONAL RINGS

There are companion results for the centroid of a circumgonal ring. For simplicity, we refer to a ring with size factor λ simply as a λ-ring. First we have:

Theorem 4.8. *The area centroid $\boldsymbol{C}(A_{\text{ring}})$ of a circumgonal λ-ring is related to the area centroid $\boldsymbol{C}(A_{\text{outer}})$ of the outer circumgon by*

$$\boldsymbol{C}(A_{\text{ring}}) = \frac{1-\lambda^3}{1-\lambda^2}\boldsymbol{C}(A_{\text{outer}}). \tag{4.9}$$

Proof. Let $\boldsymbol{C}(A_{\text{inner}})$ denote the area centroid of the inner circumgon, and let A_{outer} and A_{inner} denote the areas of the outer and inner circumgons, respectively. Equating moments we have

$$(A_{\text{outer}} - A_{\text{inner}})\boldsymbol{C}(A_{\text{ring}}) + A_{\text{inner}}\boldsymbol{C}(A_{\text{inner}}) = A_{\text{outer}}\boldsymbol{C}(A_{\text{outer}}).$$

Because of the relations

$$\boldsymbol{C}(A_{\text{inner}}) = \lambda\boldsymbol{C}(A_{\text{outer}}), \quad A_{\text{inner}} = \lambda^2\, A_{\text{outer}},$$

the foregoing equation reduces to (4.9). Now $(1-\lambda^3)/(1-\lambda^2) = (\lambda^2+\lambda+1)/(\lambda+1)$, so (4.4) is a limiting case of (4.9) as $\lambda \to 1$. In fact, our original discovery of (4.4) was obtained as this limiting case of (4.9).

Next, we extend Theorem 4.7 by relating the area centroid $\boldsymbol{C}(A_{\text{ring}})$ of a circumgonal ring with the centroid $\boldsymbol{C}(B_{\text{total}})$ of its full boundary.

Theorem 4.9. *The area centroid $\boldsymbol{C}(A_{\text{ring}})$ of a circumgonal λ-ring is related to the centroid $\boldsymbol{C}(B_{\text{total}})$ of its boundary by*

$$\boldsymbol{C}(A_{\text{ring}}) = \frac{2}{3}\frac{\lambda^2+\lambda+1}{\lambda^2+1}\boldsymbol{C}(B_{\text{total}}). \tag{4.10}$$

Proof. Denote the vectors from the incenter to the centroids of the inner and outer boundaries, respectively, by $\boldsymbol{C}(B_{\text{inner}})$ and $\boldsymbol{C}(B_{\text{outer}})$. Let P_{in} and P_{out} denote the corresponding inner and outer perimeters. The definition of centroid states that

$$\boldsymbol{C}(B_{\text{total}}) = \frac{P_{\text{in}}\,\boldsymbol{C}(B_{\text{inner}}) \;+ P_{\text{out}}\;\boldsymbol{C}(B_{\text{outer}})}{P_{\text{in}} + \;P_{\text{out}}}.$$

Because of the relation $P_{\text{in}} = \lambda P_{\text{out}}$ this becomes

$$\boldsymbol{C}(B_{\text{total}}) = \frac{\lambda\;\boldsymbol{C}(B_{\text{inner}}) \;+\; \boldsymbol{C}(B_{\text{outer}})}{\lambda \;+\; 1}. \tag{4.11}$$

But

$$\boldsymbol{C}(B_{\text{inner}}) = \lambda\boldsymbol{C}(B_{\text{outer}}),$$

and by (4.4),

$$\boldsymbol{C}(B_{\text{outer}}) = \frac{3}{2}\boldsymbol{C}(A_{\text{outer}})$$

so (4.11) can be written as

$$\boldsymbol{C}(B_{\text{total}}) = \frac{3}{2}\frac{\lambda^2+1}{\lambda+1}\boldsymbol{C}(A_{\text{outer}}),$$

which, together with (4.9), gives

$$\boldsymbol{C}(A_{\text{ring}}) = \frac{2}{3}\frac{\lambda+1}{\lambda^2+1}\frac{1-\lambda^3}{1-\lambda^2}\boldsymbol{C}(B_{\text{total}}) = \frac{2}{3}\frac{\lambda^2+\lambda+1}{\lambda^2+1}\boldsymbol{C}(B_{\text{total}}),$$

as asserted.

The next theorem relates the area centroid $\boldsymbol{C}(A_{\text{ring}})$ of a circumgonal ring to the centroids of its outer and inner boundary curves.

Theorem 4.10. *For a circumgonal λ-ring the following hold:*

$$\boldsymbol{C}(A_{\text{ring}}) = \frac{2}{3}\frac{\lambda^2+\lambda+1}{\lambda\ +1}\boldsymbol{C}(B_{\text{outer}}), \tag{4.12}$$

$$\boldsymbol{C}(A_{\text{ring}}) = \frac{2}{3}\frac{\lambda^2+\lambda+1}{\lambda(\lambda\ +\ 1)}\boldsymbol{C}(B_{\text{inner}}), \tag{4.13}$$

$$\boldsymbol{C}(A_{\text{ring}}) = \frac{2}{3}\frac{\lambda^2+\lambda+1}{1-\lambda^2}(\boldsymbol{C}(B_{\text{outer}}) - \boldsymbol{C}(B_{\text{inner}})), \tag{4.14}$$

$$\boldsymbol{C}(A_{\text{ring}}) - \boldsymbol{C}(B_{\text{inner}}) = \frac{\lambda\ +\ 2}{1+2\lambda}(\boldsymbol{C}(B_{\text{outer}}) - \boldsymbol{C}(A_{\text{ring}})). \tag{4.15}$$

Proof. Theorem 4.8 and (4.4) yield (4.12), which implies (4.13). From (4.12) and (4.13) we infer (4.14). From (4.13) we obtain

$$\begin{aligned}\boldsymbol{C}(A_{\text{ring}}) - \boldsymbol{C}(B_{\text{inner}}) &= \frac{(1-\lambda)(\lambda+2)}{3\lambda(\lambda+1)}\boldsymbol{C}(B_{\text{inner}})\\ &= \frac{(1-\lambda)(\lambda+2)}{3(\lambda+1)}\boldsymbol{C}(B_{\text{outer}}),\end{aligned}$$

whereas (4.12) gives us

$$\boldsymbol{C}(B_{\text{outer}}) - \boldsymbol{C}(A_{\text{ring}}) = \frac{(1-\lambda)(1+\ 2\lambda)}{3(\lambda+1)}\boldsymbol{C}(B_{\text{outer}}).$$

Comparing the last two equations we get (4.15).

For a trapezoid, the result in (4.15) was known to Archimedes [44, Proposition 15, p. 201]:

Corollary 4.2. (Archimedes) *The area centroid of a trapezoid lies on the segment joining the midpoints of its parallel edges and divides this segment in the ratio $(\lambda+2)/(1+2\lambda)$ when taken from the shorter parallel edge to the longer, the ratio of whose lengths is λ.*

Archimedes does not state explicitly from where the division point is measured, but this is implicit in his accompanying diagram.

4.6 EXTENSIONS TO 3-SPACE

Although it is well known that every tetrahedron circumscribes a sphere, the following two simple consequences apparently have not been previously recorded. First, a plane through the center of the inscribed sphere divides the tetrahedron into two smaller solids whose surface areas are equal if and only if their volumes are equal. Second, the centroid of the boundary surface of a tetrahedron and the centroid of its volume are always collinear with the center of the inscribed sphere, at distances in the ratio 4:3 from the center.

The rest of this chapter shows that both these and deeper results hold, not only for the tetrahedron or any polyhedron that circumscribes a sphere, but for more general solids called circumsolids (defined in Section 4.8), whose faces can be curved as well as planar. The curved faces can be cylindrical, conical, or spherical. Each circumsolid circumscribes a sphere (its insphere), and all share the following property, proved in Section 4.8:

Theorem 4.11. *The volume of a circumsolid is one-third the product of its outer surface area and the radius of its insphere.*

Section 4.7 reveals that Theorem 4.11 is implicitly contained in known formulas for volume and surface area of many familiar solids that happen to be circumsolids. Section 4.8 extends the definition of circumgon to circumsolids in 3-space. Section 4.9 applies Theorem 4.11 to interesting examples of circumsolids, including star-like circumsolids. Section 4.11 uses Theorem 4.11 to solve the difficult problem of calculating the volume of the solid of intersection of a right circular cone cut orthogonally by a right circular cylinder. The result, stated in Theorem 4.17, generalizes the classical Archimedean result for the intersection of two circular cylinders. Section 4.12 relates the volume and surface area centroids of a circumsolid. Section 4.13 determines the volume of a circumsolid shell in terms of its constant thickness, thus providing a far-reaching extension of the classical Egyptian and prismoidal formulas. Section 4.14 deals with centroids of circumsolid shells.

4.7 FAMILIAR CIRCUMSOLIDS

For a circumsolid, Theorem 4.11 tells us that $V/S = r/3$, where V is the volume, S is the outer surface area, and r is the radius of the insphere (called the inradius). The following examples illustrate this property with familiar solids that are also circumsolids.

Example 1 (Sphere). For a sphere of radius r, the known formulas $V = 4\pi r^3/3$ and $S = 4\pi r^2$ reveal that $V/S = r/3$.

Example 2 (Prism circumscribing a sphere). Not every prism circumscribes a sphere. A notable exception is a right prism whose base is a regular n-gon. The base circumscribes a circle whose radius r is that of the inscribed sphere. If each edge of the base has length a, the lateral surface area is $2nar$, and the base has area $nar/2$, so the total surface area $S = 3nar$. Its volume $V = nar^2$, which gives $V/S = r/3$.

Example 3 (Cylinder circumscribing a sphere). A right circular cylinder circumscribing a sphere of radius r has volume $V = 2\pi r^3$, lateral surface area $4\pi r^2$, and total surface area $S = 6\pi r^2$, implying that $V/S = r/3$. This also follows from Example 2, because the cylinder is a limiting case of a circumscribing prism.

Example 4 (Tetrahedron). Every tetrahedron, regular or not, circumscribes a sphere. It is known, and easy to verify, that its volume V and total surface area S are related by the formula $V = Sr/3$, where r is the inradius. In fact, if the four triangular faces have areas S_1, S_2, S_3, and S_4, then $S = S_1 + S_2 + S_3 + S_4$. On the other hand, the tetrahedron can be divided into four pyramids having a common vertex at the center of the inscribed sphere and respective volumes $V_k = S_k r/3$, whose sum is $V = Sr/3$, as asserted.

Theorem 4.11 yields a simple relation between the inradius and the four altitudes of a tetrahedron. Let h_k denote the altitude to the face of area S_k from the opposite vertex. Then $V = S_k h_k/3$, so $S_k = 3V/h_k$ for each k, and their sum is

$$S = 3V(\frac{1}{h_1} + \frac{1}{h_2} + \frac{1}{h_3} + \frac{1}{h_4}).$$

But we also have $S = 3V/r$, so

$$\frac{1}{r} = \frac{1}{h_1} + \frac{1}{h_2} + \frac{1}{h_3} + \frac{1}{h_4}.$$

In words, the reciprocal of the inradius of a tetrahedron is the sum of the reciprocals of its four altitudes. This extends a corresponding result for the inradius of a triangle: The reciprocal of the inradius of a triangle is the sum of the reciprocals of its three altitudes.

Example 5 (Pyramid circumscribing a sphere). A right pyramid whose base is a regular polygon and whose altitude passes through the center of the base circumscribes a sphere. Its volume V and total surface area S are related by $V = Sr/3$, where r is the inradius. This can be verified by dividing the pyramid into smaller pyramids with a common vertex at the incenter as was done in Example 4. In fact, the same method applies even if the polygonal base is not orthogonal to the axis of the pyramid, but slanted as shown in Figure 4.10a. Again, we find that $V = Sr/3$. For a right pyramid, this can also be derived (with more effort) from known formulas for the volume and surface area of a pyramid.

Example 6 (Cone circumscribing a sphere). As the number of edges of the polygonal base in Example 5 tends to ∞, the pyramid becomes a circular cone circumscribing the same sphere, as illustrated by Figure 4.10b. Consequently, the relation $V/S = r/3$ holds for the limiting cone. For a right circular cone, this can also be derived from known formulas expressing volume and surface area in terms of the altitude and slant height of the cone, but additional effort is required to relate these quantities to the inradius r.

The volume and surface area of a right circular cone have been studied since antiquity. Apparently no one has previously recorded the remarkable relation $V/S = r/3$ connecting its total surface area, volume, and inradius.

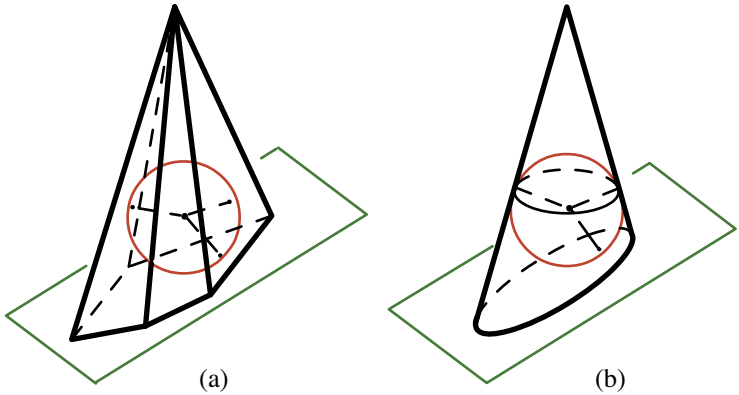


Figure 4.10: (a) Pyramid circumscribing sphere; (b) cone circumscribing sphere.

The foregoing examples illustrate the intrinsic presence of Theorem 4.11 in familiar circumsolids. To extend the applicability of Theorem 4.11 to less familiar circumsolids, we define general circumsolids in terms of simpler building blocks, by analogy with the definition of circumgons as given in Section 4.2.

4.8 BUILDING BLOCKS OF A CIRCUMSOLID

Instead of two types used in the plane we consider four types, as illustrated in Figure 4.11. This leads to a class of circumsolids with more extensive applications than the class of circumgons treated in the planar case.

Flat-faced building block.

Start with a sphere (called the insphere), and a tangent plane. In the plane, consider a region F bounded by a simple closed curve, and assume that F has a finite area. Form the union of all line segments joining the incenter to the points of F. This is a conical solid having F as its base, the incenter as its vertex, and the inradius as its altitude. The term "conical solid" is used with the understanding that the solid is a pyramid when the base is polygonal. We call this solid a flat-faced building block, and we call F its outer face (Figure 4.11a). This is the 3-dimensional analog of the triangular wedge building block in Figure 4.3a, the outer face being the analog of the outer edge. We extend the analogy further by defining the outer surface area of the building block to be the area of its outer face. Because the volume of a conical solid is one-third the area of its base times its altitude, the volume of a flat-faced building block is one-third the product of its outer surface area and its inradius.

Any polyhedral solid circumscribing a sphere is the union of a finite number of flat-faced building blocks, each of whose outer faces is polygonal. The surface area of a polyhedral solid is the sum of the areas of its polygonal faces, and its volume is the sum of the volumes of the blocks. Hence, Theorem 4.11 holds for every polyhedral circumsolid.

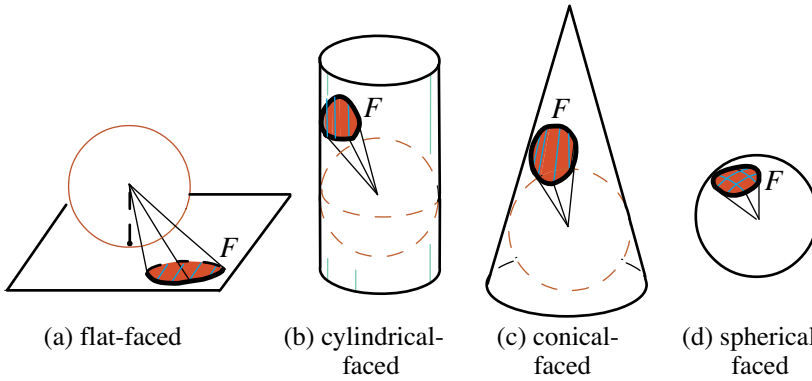


Figure 4.11: Four types of building blocks for a circumsolid.

The most general circumsolid is one that can be obtained from flat-faced building blocks by a limiting process. Again, start with a sphere (the insphere), and consider a region F of finite area lying on a surface tangent to the insphere, for example, on a developable surface with generators tangent to the insphere. The set of all line segments joining the incenter to points of F can be regarded as a building block with F as its outer face, and we call the area of F the outer surface area of the building block. The most general surface tangent to a sphere is not easy to visualize. For simplicity, we consider only four familiar types of surfaces tangent to the insphere: a plane, a cylinder, a cone, and the sphere itself. These are especially suitable for applications. The plane gives flat-faced building blocks, and we turn now to the other three types.

Cylindrical-faced building block.

In this type, shown in Figure 4.11b, the outer face F lies on a cylindrical surface tangent to the sphere. We call this a cylindrical-faced building block. Every such block is the limit of flat-faced building blocks, so its volume is one-third the product of its outer surface area and its inradius.

Conical-faced building block.

For this type, shown in Figure 4.11c, the region F lies on a conical surface tangent to the sphere. We refer to this type as a conical-faced building block. Every conical-faced building block is again the limit of flat-faced building blocks, ensuring that its volume is one-third the product of its outer surface area and its inradius.

Spherical-faced building block.

This is the 3-dimensional analog of a circular sector building block, and we call it a spherical-faced building block (see Figure 4.11d). In this case F lies on the surface of the insphere and plays the role of the circular arc intercepted by the circular sector in Figure 4.3b. As with the previous types, each spherical building block is

the limit of flat-faced building blocks. Its volume is thus one-third the product of its outer surface area and its inradius.

Definition of circumsolid. A circumsolid is the union of a finite set of nonoverlapping building blocks having the same insphere. The sum of the areas of the outer faces is called the outer surface area of the circumsolid.

This definition leads to the following equivalent formulation of Theorem 4.11: *The ratio of volume to outer surface area of any circumsolid is one-third its inradius.*

Extensions to n-space.

We can start with an n-sphere as insphere and introduce n-circumsolids by using building blocks employing portions of $(n-1)$-dimensional hyperplanes, cylinders, cones, and spherical surfaces by analogy with those shown in Figure 4.11. As might be expected, more types of building blocks are possible in higher-dimensional space. In Chapter 13 we will encounter several new types of n-circumsolids, among which are the n-cylindroid, n-double conoid, and n-hexaconoid. All the circumsolids share a fundamental property that extends Theorem 4.11 and relates their volumes and outer surface areas. If an n-circumsolid with insphere of radius r has volume V_n and outer surface area S_n, then we have:

Theorem 4.12. (Circumsolid Property) *The ratio of volume to outer surface area of an n-circumsolid is one-nth its inradius. Equivalently, we have*

$$V_n = \frac{r}{n} S_n. \tag{4.16}$$

When $n = 2$ the circumsolid is a circumgonal region with the insphere being the incircle, and Theorem 4.12 becomes Theorem 4.4. When $n = 3$, Theorem 4.12 reduces to Theorem 4.11. The following consequence of Theorem 4.12 does not depend explicitly on the inradius or the dimensionality of the space.

Theorem 4.13. *For two n-circumsolids with the same inradius, the ratio of their volumes is equal to the ratio of their outer surface areas.*

Proof. Take two n-circumsolids with the same inradius that have volumes V and V' and respective outer surface areas S and S'. From (4.16) we see that $V/V' = S/S'$, which immediately gives Theorem 4.13.

4.9 APPLICATIONS OF THEOREM 4.13

In this section we assume $n = 2$ or $n = 3$. A striking consequence of Theorem 4.13 is obtained by cutting an n-circumsolid with an $(n-1)$-dimensional plane through its incenter. (We use the term "plane" rather than the more precise term "hyperplane," with the understanding that if $n = 3$ it means an ordinary plane, while if $n = 2$ it refers to a line in the plane of the circumgon.) A finite set of planes passing through the incenter divides the solid into smaller circumsolids that have the same insphere but do not include the common $(n-1)$-dimensional faces in the dividing planes. Each pair of circumsolids has the same inradius, so Theorem 4.13 has the consequence:

Theorem 4.14. *A finite set of planes passing through the incenter of an n-circumsolid divides the circumsolid into smaller n-circumsolids in such a way that the ratio of the outer surface areas of any two is equal to the corresponding ratio of their volumes.*

Corollary 4.3. *A plane through the incenter of an n-circumsolid bisects the outer surface area if and only if it bisects the volume. In this case, the two solids so formed (including the common face, if there is one) have equal total surface areas and equal volumes.*

For a tetrahedron, this gives the result mentioned in the opening paragraph of Section 4.6, which we believe is new. The bisecting property for the case of a planar triangle is known [68]. Our simple proof, based on Theorem 4.13, makes the result seem almost trivial, not only for a triangle but for any circumgon.

The importance of Theorem 4.12 has been revealed by its connection with many familiar circumsolids and by its consequences in Theorems 4.13 and 4.14. But its real power comes into play when applied to a broader class of complicated circumsolids. For example, we can construct new families of circumsolids by replacing the regular polygonal bases of the prisms and pyramids in Examples 2 and 5 with circumgons, as suggested by the examples in Figure 4.12a and 4.12b. And we can form even more new families by truncating a circumsolid by one or more planes tangent to the insphere, as indicated by the examples in Figure 4.12c, d, and e.

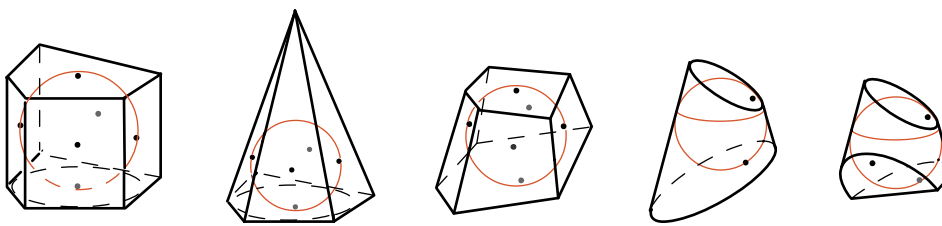


Figure 4.12: Further examples of circumsolids.

Our results can be used to analyze various intersections of circumsolids having the same insphere (see Section 4.11).

Example 7 (Archimedean dome and circumscribing prism). Figure 4.13a shows an Archimedean dome (defined in Chapter 5). This circumsolid is the union of cylindrical-faced building blocks. The axes of the cylinders are coplanar and intersect at one point. The cross section through the sphere's equator is a polygon circumscribing the equator, but the equatorial base is not one of the outer faces when the dome is regarded as a circumsolid. Each cross section by a plane parallel to the equator is a similar polygon circumscribing the circular cross section of the sphere. Figure 4.13b shows the dome circumscribed by a prism, a circumsolid with the same insphere. Cross sections of the prism parallel to the equator are congruent to the equatorial polygon in Figure 4.13a.

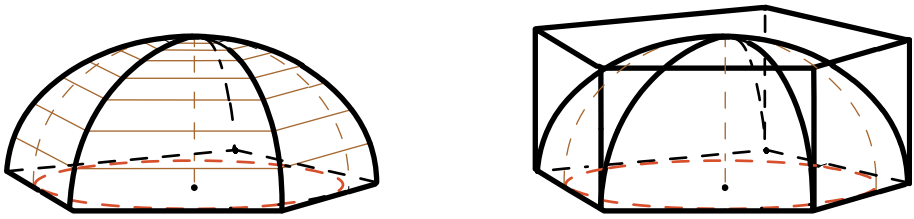


Figure 4.13: An Archimedean dome (a), and its circumscribing prismatic container (b). Neither includes the base in surface area comparisons.

Archimedean domes and their circumscribing prismatic containers are discussed in greater detail in Chapter 5, where it is shown that the ratio V_d/V_p of the dome volume V_d to the prism volume V_p is $2/3$, and that the same is true for the ratio S_d/S_p of the lateral dome area S_d to the outer prism area S_p (lateral area plus the area of the top face). This implies $V_d/S_d = V_p/S_p$, and by Theorem 4.11 this common ratio is $r/3$, where r is the inradius.

Example 8 (Star-shaped circumsolids). Like circumgons, circumsolids need not be convex, and they include star-shaped figures like stellated polyhedra obtained by extending the faces of regular convex polyhedra. Figure 4.14a shows a stellated dodecahedron formed by extending the edges of each pentagonal plane face of a regular dodecahedron until they form a pentagram. The stellated dodecahedron formed by the twelve intersecting pentagrams can also be constructed by adding twelve pyramids to the faces as indicated in Figure 4.14b. This is a circumsolid because each plane face is tangent to the insphere. More examples of star-shaped circumsolids can be formed by extending the outer plane faces of a convex polyhedral circumsolid or, for example, by suitably attaching cones to a given insphere. In each case, the ratio of volume to total surface area is one-third the inradius.

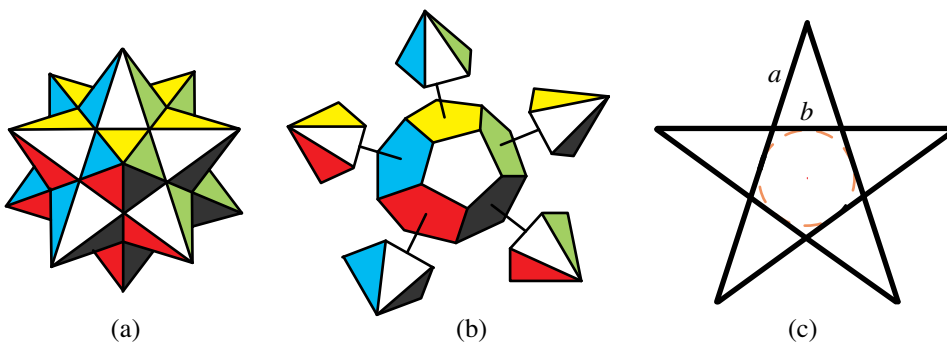


Figure 4.14: (a) Stellated dodecahedron. (b) Pyramids added to a dodecahedron. (c) Pentagram and pentagon obtained by unfolding a pyramid in (b).

A stellated polyhedron and its core polyhedron both circumscribe the same sphere. By Theorem 4.12, the ratio of their volumes is the same as the ratio of their total surface areas. The ratio can be expressed in terms of planar ratios.

To illustrate with an example, refer to Figure 4.14c, which shows a pentagram formed by unfolding the five triangular faces of a pyramid in Figure 4.14b onto a plane surrounding a regular pentagon. The pentagram has ten edges, each of length a, say, and the pentagon has five edges, each of length b. If $\gamma = a/b$, the ratio of their perimeters is $10a/(5b) = 2\gamma$. Because the pentagram and pentagon circumscribe the same incircle, the ratio of their areas is also 2γ.

The surface area of the stellated dodecahedron is equal to the total lateral surface area of the twelve pyramids, each of which consists of five congruent triangles with total area equal to that of the pentagram, minus the area of the pentagonal base. Consequently, the ratio of the surface area of the stellated dodecahedron to that of the dodecahedron is $2\gamma - 1$, which is also the ratio of their volumes. It is known that $\gamma = (1 + \sqrt{5})/2$ (the golden ratio) so $2\gamma - 1 = \sqrt{5}$.

The simple calculations of Example 8 lead to the following observations:

$$\frac{\text{area of pentagram}}{\text{area of pentagon}} = \frac{\text{perimeter of pentagram}}{\text{perimeter of pentagon}} = 2\gamma = 1 + \sqrt{5},$$

$$\frac{\text{volume of stellated dodecahedron}}{\text{volume of dodecahedron}} = \frac{\text{surface area of stellated dodecahedron}}{\text{surface area of dodecahedron}}$$
$$= 2\gamma - 1 = \sqrt{5}.$$

The use of Theorem 4.11 enabled us to deduce the volume and surface area relations by analyzing planar regions only.

4.10 OPTIMAL CIRCUMGONS AND CIRCUMSOLIDS

Two figures in the plane (or in 3-space) are said to have the same shape if one of them can be scaled to become congruent to the other. Thus, two similar figures have the same shape. A typical isoperimetric problem in the plane compares two different shapes with equal perimeters and asks for the shape with larger area, which is called optimal. In 3-space, two solids with different shapes having a given surface area are compared, and the shape with larger volume is called optimal. This section treats optimal circumgons and circumsolids. The results are of particular interest because the figures need not be convex and our comparisons lead to quantitative relations as illustrated in Examples 9 and 10.

The basic formula (4.16) for any n-circumsolid is invariant under scaling, so the size of a figure plays no role in optimality considerations. Applying (4.16) to two n-circumsolids with volumes V and V', outer surface areas S and S', and inradii r and r', respectively, we find that

$$\frac{V}{V'} = \frac{S}{S'}\frac{r}{r'},$$

from which we deduce:

Theorem 4.15. *Suppose that two different n-circumsolids have equal outer surface areas. If the ratio of their inradii is ρ, then the ratio of their volumes is also ρ. The one with the larger inradius has larger volume, hence is optimal.*

Corollary 4.4. *Among all n-circumsolids with a given outer surface area, the n-sphere has the largest inradius and the largest volume, and hence is optimal.*

Theorem 4.15 not only establishes optimality, which is a qualitative result, but it also gives a quantitative comparison of areas and volumes. This is demonstrated by the next two examples.

Example 9 (Equilateral triangle and hexagon). Figure 4.15a shows a regular hexagon and an equilateral triangle having equal perimeters. Because the hexagon is more circular than the triangle, it has larger area. What is the exact ratio of their areas? The two-dimensional version of Theorem 4.15 answers this question. The ratio of their areas is equal to the ratio of their inradii, so we need to determine the ratio of the inradii.

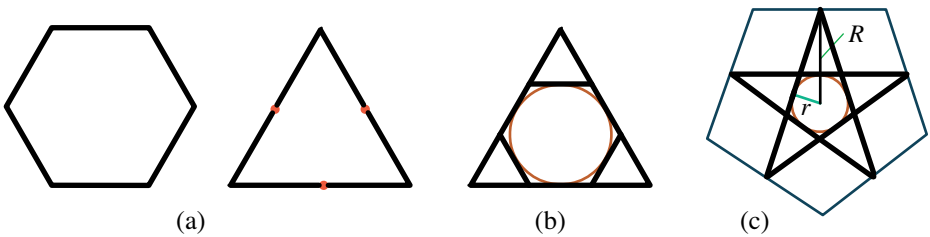


Figure 4.15: (a) Isoperimetric hexagon and triangle; (b) hexagon inscribed in triangle; (c) Example 10.

One way to do this is illustrated in Figure 4.15b, which shows a smaller regular hexagon inscribed in the equilateral triangle. Both are circumgons with the same incircle, and it is clear from Figure 4.15b that the ratio of their perimeters (triangle to hexagon) is 9 to 6, or 3/2. Therefore if we scale the inscribed hexagon from its incenter by the factor 3/2, we obtain the larger hexagon in Figure 4.15a having the same perimeter as the triangle. The larger hexagon and the original triangle are circumgons whose inradii have ratio 3/2, so by Theorem 4.15 the ratio of their areas is also 3/2. In other words, a regular hexagon with the same perimeter as an equilateral triangle has 3/2 the triangle's area. This can also be seen directly by dissecting each polygon in Figure 4.15a into smaller equilateral triangles.

Example 10 (Pentagram and regular pentagon). Suppose that a pentagram and a regular pentagon have equal perimeters, as in Figure 4.15c, where the pentagon circumscribes the pentagram.

What is the exact ratio of their areas?

The midpoints of the edges of the large regular pentagon are the vertices of the pentagram, and it is clear that the large pentagon has the same perimeter P as the pentagram. The large pentagon is a circumgon with inradius R and area $PR/2$,

and the pentragram is a circumgon with inradius r and area $Pr/2$, so the ratio of their areas is R/r, the ratio of their inradii, as predicted by Theorem 4.15. From similar triangles in Figure 4.15c we obtain $R/r = a/(b/2) = 2\gamma$ (see Example 8 and Figure 4.14c), yielding an area ratio of 2γ:

$$\frac{\text{area of large pentagon}}{\text{area of pentagram}} = 2\gamma = 1 + \sqrt{5}.$$

The same result can be obtained without explicitly constructing the large pentagon. The pentagram and the smaller pentagon have the same inradius, and we found in Example 8 that the ratio of their perimeters is 2γ. Therefore, if we expand the small pentagon from the incenter by the scaling factor 2γ, we obtain a larger pentagon with the same perimeter as the pentagram. The ratio of their inradii is 2γ, hence by Theorem 4.15 the ratio of their areas is also 2γ. The use of Theorem 4.15 is preferable because it applies in cases in which it is not clear how to construct explicitly a polygon isoperimetric to another.

The next theorem refers to two n-circumsolids with the same inradius. Theorem 4.15 tells us that the ratio of their volumes is equal to the ratio of their surface areas. Let μ denote the common ratio (larger to smaller), so that $\mu \geq 1$.

Theorem 4.16. *If two different n-circumsolids have the same inradii, then the one with smaller volume (or smaller outer area) has optimal shape. Moreover, for the same outer area the optimal shape has volume exactly μ times larger than the other shape.*

Proof. Expand the smaller solid from its incenter by the scaling factor μ to match the outer area of the larger solid. Expansion increases the inradius by the factor μ. In view of Theorem 4.15, its volume is μ times larger, and its shape is optimal.

Example 11 (Stellated dodecahedron). In Example 8 we showed that the volume of the stellated dodecahedron is $\sqrt{5}$ times that of the dodecahedron with the same insphere. Therefore, according to Theorem 4.16 with $\mu = \sqrt{5}$, for the same outer surface area a dodecahedron has $\sqrt{5}$ times larger volume than a stellated dodecahedron. Similarly, the results of Examples 9 and 10 follow immediately from Theorem 4.16.

4.11 INTERSECTION OF A CONE AND A CYLINDER HAVING THE SAME INSPHERE

Figure 4.16a shows a circular cone with a vertical axis and a circular cylinder with a horizontal axis circumscribing the same insphere of radius R.

Problem: *Find the volume of the solid of intersection of this cone and cylinder.*

The corresponding problem for two intersecting cylinders originated with Archimedes and has become a standard calculus exercise that is relatively easy to solve because all cross sections in one direction are squares. The result (see Theorem 5.3) is two-thirds the volume of its circumscribing cube, or $16R^3/3$.

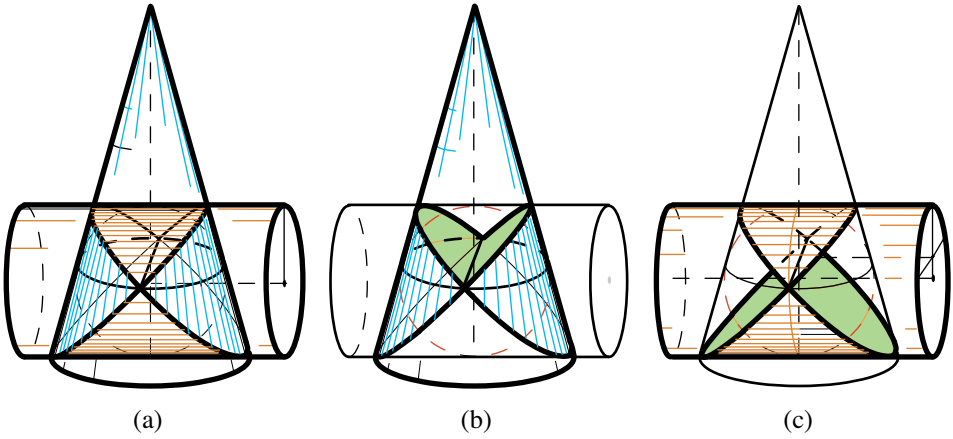


Figure 4.16: Solid of intersection (a), bounded by a conical surface (b), and a cylindrical surface (c).

Calculating the volume of the solid of intersection of a cone and cylinder is much more difficult. In Figure 4.16a, each horizontal cross section perpendicular to the axis of the cone is a rectangle capped by circular segments on two opposite edges. The cross-sectional area can be expressed as a complicated function of the vertical distance from the center of the insphere, and the volume V of the solid of intersection can be expressed as an integral of the cross-sectional areas. This approach leads to some unattractive integrals that are not easy to evaluate. They can be avoided by using Theorem 4.11. The solid in question is a circumsolid, and Theorem 4.11 tells us that $V = RS/3$, where S is the area of its outer boundary surface. This reduces the problem to that of calculating S.

The boundary surface in Figure 4.16a consists of two parts, a conical portion (Figure 4.16b) whose area we call S_1, and a cylindrical portion (Figure 4.16c) whose area we call S_2, each of which can be calculated separately without integration. We describe the calculations in detail because they involve geometric properties of the boundary surface that are of independent interest. The final results are:

Theorem 4.17. *The intersection of a right circular cone with vertex angle 2α, and an orthogonal circular cylinder circumscribing the same insphere of radius R, is a circumsolid. Its boundary surface has a conical portion of area S_1 given by*

$$S_1 = 4R^2(1 + \frac{2\alpha}{\sin 2\alpha}), \tag{4.17}$$

and a cylindrical portion of area S_2 given by

$$S_2 = 4R^2(2 + 2\alpha \tan \alpha). \tag{4.18}$$

The volume of the solid of intersection is $V = R(S_1 + S_2)/3$, or

$$V = \frac{4}{3}R^3(3 + 2\alpha \tan \alpha + \frac{2\alpha}{\sin 2\alpha}). \tag{4.19}$$

If we keep R fixed in (4.19) and let $\alpha \to 0$, the cone becomes a cylinder of radius R and the solid becomes the intersection of two orthogonal cylinders (an Archimedean globe) with volume $16R^3/3$, as expected.

The proof of Theorem 4.17 depends on a sequence of lemmas. The first, whose proof is left to the reader, deals with known properties of an ellipse, and does not involve parameters associated with the cone.

Lemma 4.1. (a) *An ellipse with semiaxes of lengths A and B (where $B \le A$) has eccentricity $\sqrt{1-(B/A)^2}$.*

(b) *If the ellipse lies on a plane inclined at angle β from the horizontal, then its projection onto a vertical plane through its minor axis is an ellipse with semiaxes of lengths B and $A\sin\beta$.*

(c) *The projected ellipse in* (b) *is a circle if and only if $B/A = \sin\beta$, in which case the inclined ellipse has eccentricity $\cos\beta$.*

The next lemma refers to an ellipse of eccentricity $\cos\beta$ cut from a right circular cone with vertex angle 2α by a plane inclined at angle β from its horizontal base, as shown in Figure 4.16. Because the eccentricity is $\cos\beta$, its vertical projection is a circle, and the circular cylinder having the circle as profile intersects the cone along the ellipse and along a symmetric congruent ellipse, as depicted in Figure 4.16. The cone and cylinder are circumsolids with a common insphere whose radius we denote by R.

Figure 4.17a shows a vertical cross section of the cone and cylinder, cut by a plane through the axis of the cone. The incenter is O, and C is the center of one of the two slanted ellipses with major axis QP. (The corresponding diagram for the congruent slanted ellipse is not shown.) The length of CO is denoted by c, written as: $|CO| = c$. When the cone and cylinder circumscribe the same insphere, α and β are related in a manner described by the next lemma, which also expresses the lengths of the axes of the ellipse in terms of the inradius R and $\sin\beta$.

Lemma 4.2. (a) *Angles α and β satisfy the relation*

$$\cos\alpha = \tan\beta. \tag{4.20}$$

(b) *The semiminor axis of the slanted ellipse has length R, and its semimajor axis has length $R/\sin\beta$.*

Proof. First we introduce some notation. In Figure 4.17a, M is the midpoint of the tangent segment NP, and C is the midpoint of QP. Therefore MC is parallel to NQ, so the angle OMC is equal to α. A vertical line through O intersects the base at W, and a vertical line through C intersects NP at T and QW at S. The incircle touches NQ at Z. Therefore NM and NZ have equal lengths that we denote by t, while QZ and QW have equal lengths that we denote by s. Finally, let $a = |QS| = |TP|$, and let $d = |MC|$. Now $c = |CO| = |TM|$, implying that $t = a - c$ because $t = |MP|$. Also, MC is parallel to NQ, so $2d = s + t$. But $s = a + c$ and $t = a - c$, so $2d = s + t = (a + c) + (a - c) = 2a$, whence

$$a = d. \tag{4.21}$$

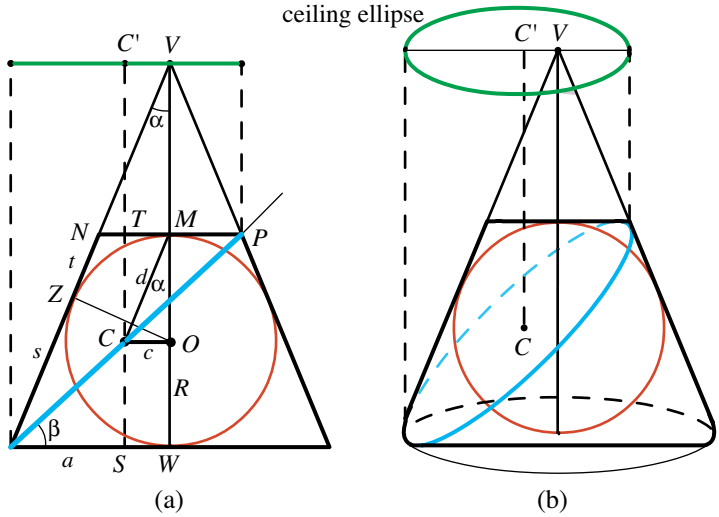


Figure 4.17: (a) Vertical cross section of the intersection of a cone and cylinder circumscribing the same insphere; (b) the ceiling ellipse is the vertical projection of the slanted ellipse onto a horizontal plane through the vertex of the cone.

Next, the right triangle COM shows that $R = d\cos\alpha$, and the right triangle QSC shows that $R = a\tan\beta$. Because $a = d$, we get (4.20), which proves part (a).

The length of the semiminor axis of the ellipse is R. A glance at the triangle QSC reveals that the semimajor axis has length $A = |QC| = R/\sin\beta$, which proves part (b).

Ceiling ellipse.

The projection of the slanted elliptical cross section of the cone onto the ceiling plane, a horizontal plane through vertex V, is called a *ceiling ellipse* (Figure 4.17b). Its semiminor axis has length R and its semimajor axis has length $a = R/\tan\beta$. The center O of the incircle projects onto the vertex V, and the center C of the ellipse projects onto a point we denote by C'.

Lemma 4.3. *The ceiling ellipse has eccentricity* $\sin\alpha$ *and one focus at* V.

Proof. First we determine the eccentricity. The ceiling ellipse has semiminor axis of length R and semimajor axis of length $a = R/\tan\beta$, so by Lemma 4.1a, its eccentricity is $\sqrt{1-\tan^2\beta}$. In light of (4.20), this is $\sqrt{1-\cos^2\alpha} = \sin\alpha$.

To prove that V is a focus, it suffices to show that

$$|C'V| = a\sin\alpha, \tag{4.22}$$

which says that $|C'V|$ is the length of the semimajor axis times the eccentricity. But this follows at once from triangle COM in Figure 4.17a and (4.21), because

$$|C'V| = c = d\sin\alpha = a\sin\alpha.$$

Lemma 4.4. *Suppose that a right pyramid has a regular n-gon as base and altitude through the incenter of the base. If α signifies the angle each face makes with the altitude, then a region of area S on the lateral surface projects onto a plane region of area $S\sin\alpha$ on the ceiling plane. The same is true of a region of area S on the lateral surface of a right circular cone with vertex angle 2α.*

Proof. Each triangular face of the lateral surface of the pyramid with area T projects onto a triangle of area $T\sin\alpha$ in the ceiling plane. Hence any subregion of the lateral surface of the pyramid with area S has a ceiling projection of area $S\sin\alpha$. This property is independent of n, and we obtain the result for a cone by letting $n \to \infty$.

Calculation of the conical portion S_1.

We refer to Figure 4.18a. In the ceiling plane the shaded region is the ceiling projection of the conical portion of area S_1 and has area equal to $S_1\sin\alpha$. This area is the sum of the areas of two confocal overlapping ellipses, minus twice the

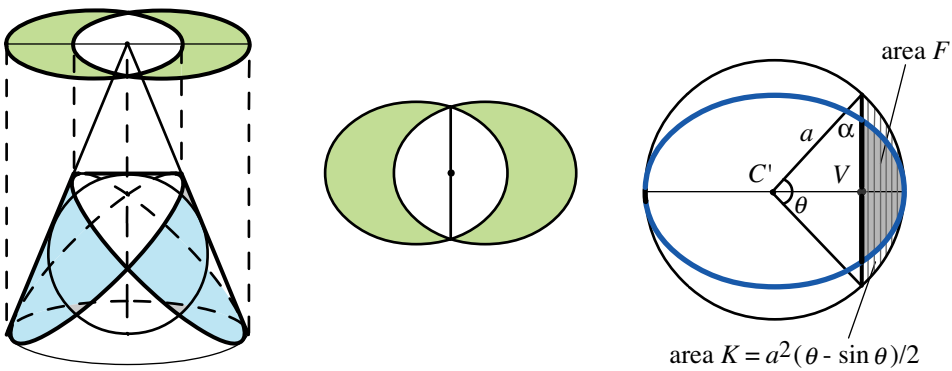


Figure 4.18: (a) Two conical wedges cut by inclined planes; (b) ceiling projection; (c) calculation of the area F of the focal segment.

area of their intersection. That is, $S_1\sin\alpha = 2(E - I)$, where E is the area of the region bounded by each ceiling ellipse and I is the area of the intersection (Figure 4.18b). The area I is equal to $2F$, where F is the area of the focal segment cut from the ellipse by a chord through its focus perpendicular to the major axis. Hence we have

$$S_1\sin\alpha = 2E - 4F. \tag{4.23}$$

Now we calculate E and F separately. Each ceiling ellipse has semiaxes of lengths R and $R/\tan\beta$, so $E = \pi R^2/\tan\beta$, and from (4.20) we find that

$$E = \frac{\pi R^2}{\cos\alpha}. \tag{4.24}$$

The next lemma evaluates F in terms of R and α.

Lemma 4.5. *The elliptical focal segment has area F given by*

$$F = R^2\left(\frac{\pi - 2\alpha}{2\cos\alpha} - \sin\alpha\right). \tag{4.25}$$

Proof. In Figure 4.18c the focal segment of area F is shown shaded. Inscribe the ceiling ellipse in a circle with center at C' and with radius equal to a, the length of the semimajor axis. Because the ellipse has eccentricity $\sin\alpha$, a radial line from C' intersects the vertical chord through the focus V at an angle α. The radial line and its mirror image subtend a central angle $\theta = \pi - 2\alpha$, and the circular segment cut by the vertical chord through V has area $K = a^2(\theta - \sin\theta)/2$. But $a = R/\cos\alpha$, which leads to

$$K = \frac{1}{2}\left(\frac{R}{\cos\alpha}\right)^2(\theta - \sin\theta) = R^2\left(\frac{\pi - 2\alpha}{2\cos^2\alpha} - \frac{\sin 2\alpha}{2\cos^2\alpha}\right) = R^2\left(\frac{\pi - 2\alpha}{2\cos^2\alpha} - \frac{\sin\alpha}{\cos\alpha}\right).$$

Now $\cos\alpha$ is the dilation factor in the vertical direction that converts the circular segment to the elliptical focal segment, so $K\cos\alpha = F$, and we obtain (4.25). Finally, use (4.25) and (4.24) in (4.23), then divide by $\sin\alpha$, to obtain (4.17) in Theorem 4.17.

Calculation of the cylindrical portion S_2.

First we recall a known property of an ellipse unwrapped from a right circular cylinder, as discussed in Section 5.8. Cut a right circular cylinder of radius R by a plane through a diameter of its base at an angle of inclination γ, where $0 < \gamma < \pi/2$. The example in Figure 4.19a shows part of the elliptical cross section and a wedge

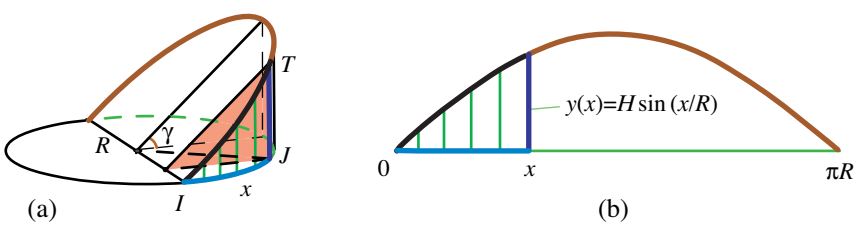


Figure 4.19: The circular arc IJ of length x in (a) unwraps onto the line segment $[0,x]$ in (b); the altitude of triangle T unwraps onto the height given by $y(x) = H\sin(x/R)$, where $H = R\tan\gamma$.

cut from the cylinder. A vertical cutting plane parallel to the major axis of the ellipse intersects the wedge along a right triangle T (shown shaded), with base angle γ. When the surface of the cylinder is unwrapped onto a plane, the circular base unfolds along a line we call the x-axis. Here x is the length of the circular arc measured from I at the extremity of the base diameter to J at the base of the triangle T, as shown in Figure 4.19a. The base of T has length $R\sin(x/R)$, and its

height is given by $H\sin(x/R)$, where $H = R\tan\gamma$. Therefore the unwrapped curve is the graph of the sinusoidal function y with period $2\pi R$ and amplitude H given by $y(x) = H\sin(x/R)$.

To adapt the same idea to the solid in Figure 4.16, tip the solid so that the cylinder's axis is vertical, as in Figure 4.20. The cutting plane is inclined at an

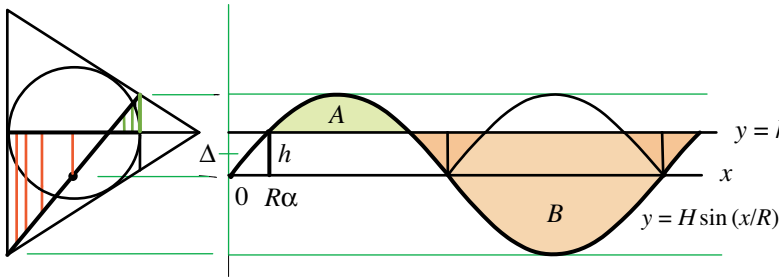


Figure 4.20: The lateral surface area S_2 is twice the sum $A+B$ of the areas of the shaded sinusoidal regions.

angle $\gamma = \pi/2 - \beta$ with a horizontal circular cross section of the cylinder and passes through a chord of the circle at vertical distance h, say, from the diameter. It cuts the lateral surface of the cylinder into two portions, the sum of whose areas is half the area S_2 in question. When unwrapped onto a plane, they form two shaded regions bounded by the curve $y = H\sin(x/R)$ and the horizontal line $y = h$. The area of the shaded region below the sine curve and above the line $y = h$ is denoted by A, while that of the region below the line $y = h$ and above the sine curve is denoted by B. The lateral surface area S_2 is $2(A + B)$.

Twice the area of the region under one arch of the curve $y = H\sin(x/R)$ is $4HR$, and Figure 4.20 shows that $4HR = A + B - 4\Delta$, where Δ is the area of the curvilinear triangular region above the portion of the sine curve over the interval $[0, R\alpha]$ and below the line $y = h$. The distance $R\alpha$ is the length of the unwrapped

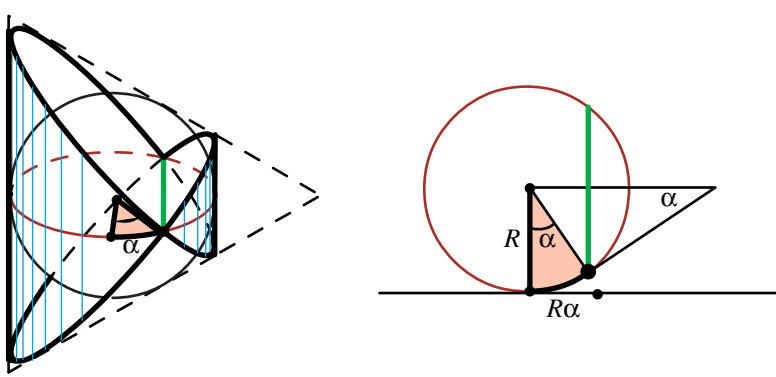


Figure 4.21: The unwrapped circular arc of radius R subtends angle α; its length is $R\alpha$.

circular arc of radius R subtending angle α (see Figure 4.21). The area Δ, in turn, is the area of a rectangle of base $R\alpha$ and altitude h, minus the area of the curvilinear triangular region below the sine curve. Because $h = H\sin\alpha$ we discover (by Corollary 5.4 in Chapter 5) that

$$\Delta = R\alpha h - HR(1-\cos\alpha),$$

hence

$$A + B = 4HR + 4\Delta = 4RH(\alpha\sin\alpha + \cos\alpha).$$

Now $H = R\tan\gamma = R\tan(\pi/2-\beta) = R\cot\beta = R/\cos\alpha$ as a result of (4.20). Therefore,

$$A + B = 4R^2(\alpha\tan\alpha + 1),$$

which implies (4.18) and completes the proof of Theorem 4.17.

4.12 CENTROIDS OF CIRCUMSOLIDS

In Theorem 4.7 we described a simple but surprising connection between the area centroid of a circumgon and the centroid of its boundary: they are collinear with the incenter, at distances in the ratio 3:2 from the incenter. Now we deduce a corresponding result for circumsolids. Specifically, denote by $\boldsymbol{C}(S)$ the vector from the incenter O to the centroid of its outer boundary surface and by $\boldsymbol{C}(V)$ the vector from O to the volume centroid of the solid. The location of one of the centroids determines the location of the other. In fact, we have:

Theorem 4.18. *The volume centroid $\boldsymbol{C}(V)$ of a circumsolid and the centroid $\boldsymbol{C}(S)$ of its outer boundary surface are collinear with the incenter and are related by*

$$\boldsymbol{C}(S) = \frac{4}{3}\boldsymbol{C}(V). \tag{4.26}$$

Proof. Consider first a flat-faced pyramidal building block whose outer face is a triangle. The centroid $\boldsymbol{C}(S)$ of the outer triangular face is in the plane of the triangle (at the intersection of the three medians). The centroid $\boldsymbol{C}(V)$ of the pyramid with the outer face as base has its centroid at a distance three-fourths the distance from the vertex to the centroid of the base. Because $\boldsymbol{C}(V)$ lies on the line segment joining the incenter to $\boldsymbol{C}(S)$ we have $\boldsymbol{C}(V) = (3/4)\boldsymbol{C}(S)$, which implies (4.26) for each pyramidal building block with a triangular outer face.

Now take a flat-faced circumsolid with polygonal outer faces and divide each outer face into triangular regions having outer areas $S_1,\ldots,S_n$, a common vertex at the incenter O, and respective volumes $V_1,\ldots,V_n$. Denote by $\boldsymbol{C}(V_1),\ldots,\boldsymbol{C}(V_n)$ the corresponding vectors from the incenter O to the volume centroid of each pyramidal block with a triangular outer face. The vectors from O to the volume centroid $\boldsymbol{C}(V)$ of their union and to the area centroid $\boldsymbol{C}(S)$ of the boundary are given by

$$\boldsymbol{C}(V) = \frac{\sum_{k=1}^n V_k\,\boldsymbol{C}(V_k)}{\sum_{k=1}^n V_k},\quad \boldsymbol{C}(S) = \frac{\sum_{k=1}^n S_k\,\boldsymbol{C}(S_k)}{\sum_{k=1}^n S_k}. \tag{4.27}$$

In the first fraction, use $V_k = S_k\, r/3$, where r is the inradius, and in the second fraction apply (4.26) to each pyramidal block to find $\boldsymbol{C}(S_k) = (4/3)\boldsymbol{C}(V_k)$. Then (4.27) implies (4.26) for a polyhedral circumsolid. Because the other types of building blocks can be regarded as limiting cases of polyhedral circumsolids, (4.26) also holds for all four types of building blocks and hence for all circumsolids.

Theorem 4.18 was stated for a tetrahedron in the opening paragraph of Section 4.6. It extends Theorem 4.7, a corresponding result for circumgons, that contains the fraction 3/2 in place of 4/3.

Example 12 (Archimedean dome and circumscribing prism). Recall the Archimedean dome and its circumscribing prism shown in Figure 4.13. In Chapter 5, Corollary 5.5, it is shown that the centroid of the surface of an Archimedean dome is at the midpoint of its altitude. If the dome is regarded as a circumsolid with inradius r, its outer surface area centroid is at distance $r/2$ from the base. Therefore, by Theorem 4.18, the volume centroid of the dome is at distance $3r/8$ from the base. This generalizes a result found by Archimedes for a hemisphere. Actually, the result for the volume centroid also holds for each component wedge of the Archimedean dome. Moreover, if each wedge of inradius r is dilated vertically by a factor λ to form a semielliptical wedge of altitude $h = \lambda r$, the elongated wedge is no longer a circumsolid, but its volume centroid is scaled by the same factor λ to a distance $3h/8$ from the horizontal base.

The circumscribing prism in Figure 4.13b has its volume centroid at distance $r/2$ from the base. Hence, by Theorem 4.18, the outer surface area of this circumsolid (lateral surface area plus the top face) has its centroid at a distance $2r/3$ from the base.

Example 13 (Right circular cone). The volume centroid of a right circular cone of altitude h is known to be on its axis at a distance $h/4$ from the cone's base. If the inradius is r, the volume centroid is at a distance $r - h/4$ from the incenter. According to Theorem 4.18, the area centroid of the total surface of the cone is at a distance $4(r - h/4)/3$ from the incenter, hence at height $(h - r)/3$ from the base of the cone. In other words, the height of the centroid of the total surface area of a cone above its base equals one-third the distance $h - r$ between the incenter and the vertex. The same is true for any right pyramid with a circumgonal base and altitude passing through the incenter of the base.

4.13 CIRCUMSOLID SHELLS

Circumsolid shells are analogous to the circumgonal rings in the plane introduced in Section 4.3. For a solid Q and a scalar λ with $0 < \lambda < 1$, choose a point O in Q and let λQ denote the solid consisting of all points $\lambda \boldsymbol{q}$ scaled from O as $\boldsymbol{q}$ ranges through all points of Q, and let Q_λ denote the solid shell lying between λQ and Q. The prototype is a spherical shell between two concentric spheres of radii λr and r.

We are interested in the case in which Q is a circumsolid with incenter O and inradius r. Then λQ is also a circumsolid with the same incenter and with inradius λr. The inner and outer boundary surfaces of a circumsolid shell Q_λ are parallel,

in the sense that the perpendicular distance between them is a constant, equal to $(1-\lambda)r$, where r is the inradius of the larger circumsolid Q. This constant, which we denote by w, is called the thickness of the shell:

$$w = (1-\lambda)r. \tag{4.28}$$

This proves part (a) of the following theorem:

Theorem 4.19. (a) *Every circumsolid shell has constant thickness.*

(b) *Conversely, any solid shell with constant thickness lying between two similar solids λQ and Q is necessarily a circumsolid shell.*

Theorem 4.19 extends Theorem 4.5 for circumgonal rings. Part (b) can be proved by passing a plane through the center of similarity perpendicular to each pair of faces, thus reducing the problem to the 2-dimensional case. We omit details.

Volume-surface area relations for circumsolid shells.

If the outer solid Q of a shell has area S and volume V, the inner solid λQ has area $T = \lambda^2 S$ and volume $\lambda^3 V$. The shell Q_λ itself has total surface area $S' = S + T = (1+\lambda^2)S$, and volume V', where

$$V' = (1-\lambda^3)V = (1-\lambda)(1+\lambda+\lambda^2)V.$$

If the outer solid is a circumsolid with volume V, then $V = Sr/3$, from which we find, using (4.28). that

$$V' = \frac{r}{3}(1-\lambda)(1+\lambda+\lambda^2)S = \frac{w}{3}(S+\lambda S+\lambda^2 S).$$

But $\lambda^2 S = T$, so $\lambda = \sqrt{T/S}$, and the formula for V' becomes

$$V' = \frac{w}{3}(S+\sqrt{ST}+T). \tag{4.29}$$

This has the same form as the famous Egyptian formula for the volume of a truncated square pyramid, where S and T are areas of the square planar bases. In our version of (4.29), the inner and outer faces need not be planar. It can be used, for example, to calculate the volume of a hemisphere, taken as a circumsolid shell with inner surface area $T = 0$. In (4.29), the term $\sqrt{ST}$, the geometric mean of S and T, is called the *mixed area* of S and T. The quantity $(S+\sqrt{ST}+T)/3$ that multiplies w is the average of S, T, and the mixed area $\sqrt{ST}$. We give it its own name:

Definition. We call the quantity $(S+\sqrt{ST}+T)/3$ the *mixed average surface area* of the circumsolid shell and denote it by S_{ave}.

Thus, (4.29) reduces to $V' = wS_{\text{ave}}$, and gives us:

Theorem 4.20. *The volume of a circumsolid shell is the product of its mixed average surface area and its thickness.*

Theorem 4.20 extends Theorem 4.6, which states the area of a circumgonal ring is one-half the product of its total perimeter and its width.

It is not easy to interpret the mixed area term $\sqrt{ST}$ in (4.29) geometrically in terms of areas. But (4.29) can be written as (4.30), in which all terms refer to areas. This alternative form involves the area of the surface midway between the outer and inner surfaces, that is, the surface whose inradius is $r(1+\lambda)/2$, the average of λr and r. We call the area of this midway surface the *midway area* and denote it by $S_{1/2}$. The midway surface is the outer surface scaled by $(1+\lambda)/2$, hence $S_{1/2} = (1+\lambda)^2 S/4$, so

$$4S_{1/2} = (1+\lambda)^2 S = S + 2\lambda S + \lambda^2 S = S + 2\sqrt{ST} + T,$$

from which we infer that

$$4S_{1/2} + S + T = 2(S + \sqrt{ST} + T) = 6S_{\text{ave}}.$$

Now we use (4.29) to obtain the following extension of the classical prismoidal formula.

Theorem 4.21. *A circumsolid shell of thickness w, outer area S, inner area T, and midway area $S_{1/2}$ has volume V' given by*

$$V' = \frac{w}{6}(S + 4S_{1/2} + T). \tag{4.30}$$

In (4.30), the term multiplying w is a weighted arithmetic mean of areas S, T, and $S_{1/2}$. When the circumsolid shell is flat-faced with parallel outer planar faces at distance w apart, (4.30) becomes the classical prismoidal formula.

Because of Theorem 4.19b, circumsolid shells are the most general shells with curved surfaces for which both the Egyptian formula and the prismoidal formula give the exact volume. For example, the volume of a hemisphere satisfies both (4.29) and (4.30) when it is considered as a circumsolid shell (with inner surface area $T = 0$). By contrast, with standard use of planar cross sections, the prismoidal formula gives the correct answer, but the Egyptian formula does not. Another example is a general solid angle or a truncated version.

4.14 CENTROIDS OF CIRCUMSOLID SHELLS

A companion result to Theorem 4.18 relates the volume centroid of the circumsolid shell Q_λ to that of the outer circumsolid Q. It extends Theorem 4.8, a corresponding result for circumgonal rings.

Theorem 4.22. *The volume centroid $\boldsymbol{C}(Q_\lambda)$ of the circumsolid shell Q_λ is related to the volume centroid $\boldsymbol{C}(Q)$ of the outer circumsolid Q by*

$$\boldsymbol{C}(Q_\lambda) = \frac{1-\lambda^4}{1-\lambda^3}\boldsymbol{C}(Q). \tag{4.31}$$

Proof. The volume of Q_λ is $V(Q) - V(\lambda Q)$. Equating moments we have

$$(V(Q) - V(\lambda Q))\boldsymbol{C}(Q_\lambda) + V(\lambda Q)\boldsymbol{C}(\lambda Q) = V(Q)\boldsymbol{C}(Q). \tag{4.32}$$

But $V(\lambda Q) = \lambda^3 V(Q)$, $V(Q) - V(\lambda Q) = V(Q)(1 - \lambda^3)$, and $\boldsymbol{C}(\lambda Q) = \lambda \boldsymbol{C}(Q)$, so (4.32) becomes

$$(1 - \lambda^3)\boldsymbol{C}(Q_\lambda)V(Q) = (1 - \lambda^4)\boldsymbol{C}(Q)V(Q),$$

which implies (4.31).

Theorem 4.18 can be obtained as a limiting case of (4.31) as $\lambda \to 1$ because the shell Q_λ has constant thickness and we can regard the area centroid of the boundary surface of Q as the limiting case of the volume centroid of the shell Q_λ as $\lambda \to 1$.

The following extension of Theorem 4.22 relates the volume centroid of the shell Q_λ with the area centroid of its full boundary S_λ (inner plus outer):

Theorem 4.23. *The volume centroid $\boldsymbol{C}(Q_\lambda)$ of a circumgonal shell Q_λ is related to its area centroid $\boldsymbol{C}(S_\lambda)$ of its full boundary by*

$$\boldsymbol{C}(Q_\lambda) = \frac{3}{4}\frac{1 - \lambda^4}{1 - \lambda^3}\frac{1 + \lambda^2}{1 + \lambda^3}\boldsymbol{C}(S_\lambda).$$

This extends Theorem 4.9, which is the planar version. We omit the proof. We also note the n-dimensional version of Theorem 4.7 relating centroids of circumsolids in n-space:

$$\boldsymbol{C}(S_n) = \frac{n+1}{n}\boldsymbol{C}(V_n). \tag{4.33}$$

NOTES ON CHAPTER 4

The material in this chapter originally appeared in [16] and [18], the first of which was awarded a Lester R. Ford Award in August, 2005.

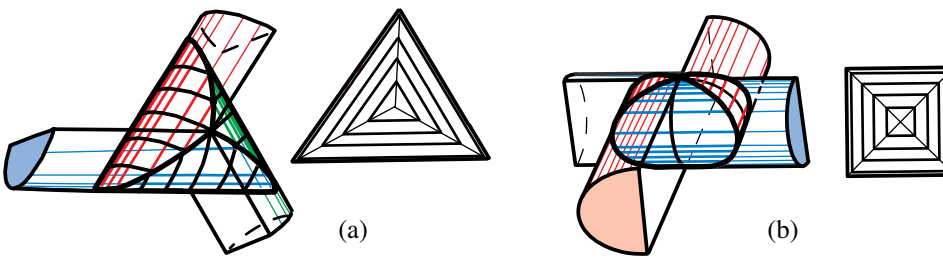
It has long been known that the centroid of the boundary of a triangle need not be at the same point as the centroid of its interior. In exploring this fact, we discovered that the two centroids are always collinear with the center of the inscribed circle, at distances in the ratio 3 : 2 from the center. The original motivation for the research leading to [16] was to generalize this elegant and surprising result to any polygon that circumscribes a circle. In doing so we found a corresponding result for any circumgon, and this led naturally to [18], which gave corresponding results in 3-space, first for any tetrahedron, and then for any circumsolid. In the process we realized that the essential key to this research was to extend Archimedes' formula for the area of a circular disk to any circumgon, and to extend his formula for the volume of a sphere to any circumsolid.

Chapter 5

THE METHOD OF PUNCTURED CONTAINERS

These problems can be easily solved by the methods developed in this chapter. The reader may wish to try solving them before reading the chapter.

In (a), three semicircular cylinders intersect to form a solid with a general triangular cross section in the plane through the three axes. The intersection of four semicircular cylinders with a square equatorial cross section is shown in (b). Slices of the solids and their inscribed spheres cut by planes parallel to the equatorial plane are also shown.



Compare corresponding slices to show that:

The total surface area S of each solid is four times the area A of the equatorial polygonal cross section. The volume V of each solid is $V = SR/3$, where R is the radius of the inscribed sphere.

CONTENTS

PART 1: ARCHIMEDEAN GLOBES

PART 2: GENERALIZED ARCHIMEDEAN DOMES

The method of punctured containers goes to the heart of some of the landmark discoveries of Archimedes concerning properties of the sphere. Part 1 of this chapter introduces the method and applies it to a special family of circumsolids that we call Archimedean globes. Cross sections of globes by planes parallel to the equatorial plane are disks bounded by similar polygons that circumscribe the circular cross sections of the sphere. Like the sphere, which is a limiting case, the volume and surface area of an Archimedean globe are two-thirds that of its circumscribing prismatic container. The results are obtained geometrically. The volume and surface area of an Archimedean shell (the region between two Archimedean globes) are also determined. Surprising consequences of these new results are: several families of incongruent solids having equal volume and equal total surface area; the quadrature of the sine curve; and the location of the centroid of any slice of a circumsolid surface.

Part 2 applies the method to more general solids, including those with nonuniform densities.

PART 1: ARCHIMEDEAN GLOBES

5.1 INTRODUCTION

A spectacular landmark in the history of mathematics was the discovery by Archimedes that the volume of a solid sphere is two-thirds the volume of the smallest right cylinder that surrounds it, and that the surface area of the sphere is also two-thirds the total surface area of the same cylinder. Archimedes was so excited by this discovery that he wanted a sphere and its circumscribing cylinder engraved on his tombstone, even though there were many other great accomplishments for which he would be forever remembered. He made this particular discovery by balancing

slices of a sphere and cone against slices of a larger cylinder, with diameter twice that of the sphere, using centroids and the principle of the lever, which were also among his remarkable discoveries.

The volume ratio for the sphere and cylinder can be derived from first principles without using levers and centroids (see [55]). This simpler and more natural method, presented in Sections 5.2 and 5.3, paves the way for generalizations. Section 5.4 introduces a family of solids circumscribing a sphere. Cross sections cut by planes parallel to the equatorial plane are disks bounded by similar n-gons that circumscribe the circular cross sections of the sphere. We call these solids *Archimedean globes* in honor of Archimedes, who treated the case $n = 4$. The sphere is a limiting case, $n \to \infty$. Each globe is analyzed by dividing it into wedges with two planar faces and one semicircular cylindrical face. In fact, Archimedes discussed (both mechanically and geometrically) volumes of wedges of this type. Figure 5.1 shows the top view of examples of globes with $n = 3$, 4, 6, and the limiting sphere.

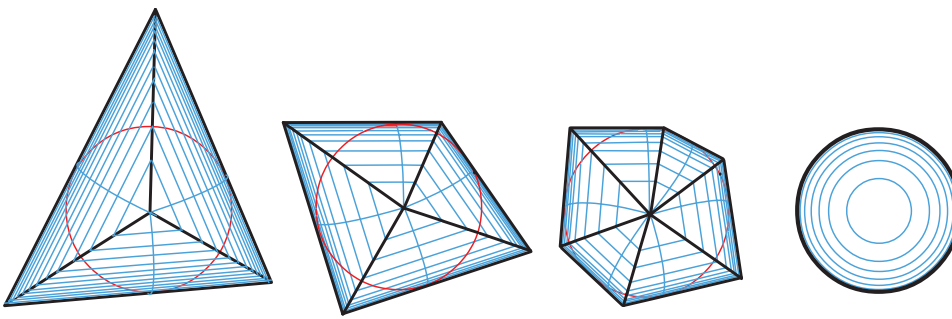


Figure 5.1: Archimedean globes (top view) showing equators for $n = 3, 4, 6$, and ∞.

As in the case of a sphere, both the volume and surface area of an Archimedean globe are two-thirds those of its circumscribing prismatic container. Section 5.5 treats the volume of an Archimedean shell, the region between two concentric Archimedean globes. The results are applied in Section 5.6 to find the surface area of an Archimedean globe. A common thread in all this work is the reduction of a problem to a simpler problem. As surprising consequences of these new results, we also obtain other families of incongruent solids having both equal volume and equal total surface area (Section 5.7), the quadrature of the sine curve (Section 5.8), and the centroid of any slice of a spherical surface (Section 5.9).

5.2 VOLUME OF A SPHERE

We present first a geometric derivation of the volume relation between a sphere and its circumscribing cylinder. By symmetry, it suffices to consider a hemisphere and its circumscribing cylinder (whose radius is equal to its altitude), as shown in Figure 5.2. A cone with the same altitude is drilled out of the center of the cylinder. The cone's volume is one-third that of the cylinder, so the solid that remains is a punctured cylinder with volume two-thirds that of the cylinder. To show that the

punctured cylinder has the same volume as the hemisphere, slice both solids by an arbitrary horizontal plane parallel to the base and note that corresponding cross sections have equal areas. Now we invoke the following:

Slicing principle. *Two solids have equal volumes if their horizontal cross sections taken at any height have equal areas.*

This statement is often called *Cavalieri's principle* in honor of Bonaventura Cavalieri (1598-1647), who attempted to prove it for general solids. Archimedes used it sixteen centuries earlier for special solids, and he credits Eudoxus and Democritus for using it even earlier in their discovery of the volume of a cone. Cavalieri employed it to find volumes of many solids, and tried to establish the principle for general solids by applying Archimedes' method of exhaustion, but it was not demonstrated rigorously until integral calculus was developed in the 17th century. We prefer using the neutral and more descriptive term *slicing principle*.

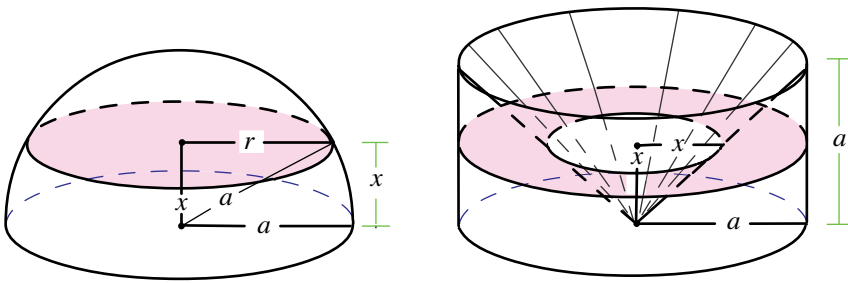


Figure 5.2: Cross sections of sphere and punctured cylinder have equal areas.

To verify the equality of the cross-sectional areas in Figure 5.2, assume that the sphere has radius a and that the cutting plane is at a distance x from the center. It cuts a circular cross section of radius r, say, with area πr^2. The corresponding cross section of the punctured cylinder is an annulus with outer radius a and inner radius x (because the altitude and radius of the cylinder are equal), so its area is equal to $\pi a^2 - \pi x^2$. But r and x are the legs of a right triangle with hypotenuse a, hence $\pi r^2 = \pi a^2 - \pi x^2$. In words, corresponding cross sections of the two solids have equal areas.

Therefore, any two planes parallel to the base cut off solids that have equal volumes. Thus, we have proved the following theorem, illustrated in Figure 5.3.

Theorem 5.1. *Any two planes parallel to the base cut the sphere and the punctured circumscribing cylinder in solid slices of equal volumes.*

Corollary 5.1. (Archimedes) *The volume of a sphere is two-thirds the volume of its circumscribing cylinder.*

Throughout this chapter we express the main theorems in the style of Archimedes, that is, by relating the volume (or surface area) of one solid to that of a simpler one. Explicit formulas in terms of dimensions of the figures can be deduced from these theorems.

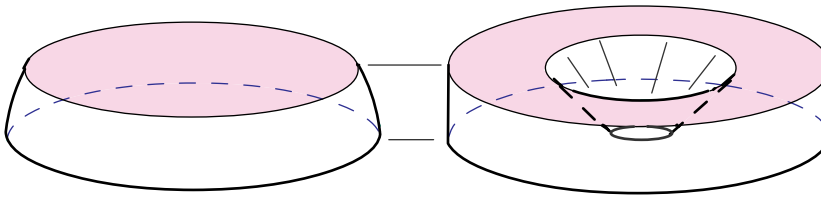


Figure 5.3: Two parallel planes cut the sphere and punctured cylinder in solid slices of equal volumes.

Examples (Formulas for volumes). If a is the radius of the hemisphere in Figure 5.2, the volume of the punctured cylinder is two-thirds the area of its base times its altitude, or $2\pi a^3/3$, which is also the volume of the hemisphere. The entire sphere has volume $4\pi a^3/3$.

Theorem 5.1 also gives the volume of a spherical segment, the portion of a sphere of radius a above a plane at a distance $R \le a$ parallel to the equatorial plane. The corresponding punctured circumscribing cylinder has altitude R, from which a cone of volume $\pi R^3/3$ has been removed. The difference $\pi a^2 R - \pi R^3/3$ is the volume of the corresponding portion of the sphere. Subtracting this from the volume of the hemisphere we get

$$\frac{2}{3}\pi a^3 + \frac{1}{3}\pi R^3 - \pi a^2 R$$

as the volume of the spherical segment. This can be written as $\pi h^2(3a - h)/3$, where $h = a - R$ is the height of the segment. Archimedes derived this by another method, but expressed it as the ratio $(a + 2R)/(2R)$ times the volume of a cone of altitude h inscribed in the spherical segment.

5.3 VOLUME OF A SPHERICAL SHELL

A spherical shell is the region between two concentric spheres. As expected, its volume is the difference of the volumes of the two spheres. A somewhat unexpected result is obtained by taking a cross section of the spherical shell by a plane that cuts both spheres. Suppose the inner and outer radii are r and a, respectively, and that the cutting plane is at a distance x from the center. We assume that $0 \le x \le r$, so the plane cuts both spheres. The cross section is an annular ring (Figure 5.4a) of outer radius s and inner radius t, say, with area $\pi s^2 - \pi t^2$. The area is independent of x: it's the same for all cutting planes. This follows from Mamikon's sweeping-tangent theorem, as already noted in Section 1.3.

We can also see it directly by applying the Pythagorean Theorem twice, once to a right triangle with hypotenuse a, and again to a right triangle with hypotenuse r (see Figure 5.4a). This gives

$$s^2 = a^2 - x^2, \quad t^2 = r^2 - x^2,$$

hence $\pi(s^2 - t^2) = \pi(a^2 - r^2)$. Therefore the area of each annular ring is equal to $\pi a^2 - \pi r^2$, which is independent of x. It is also the cross-sectional area of the

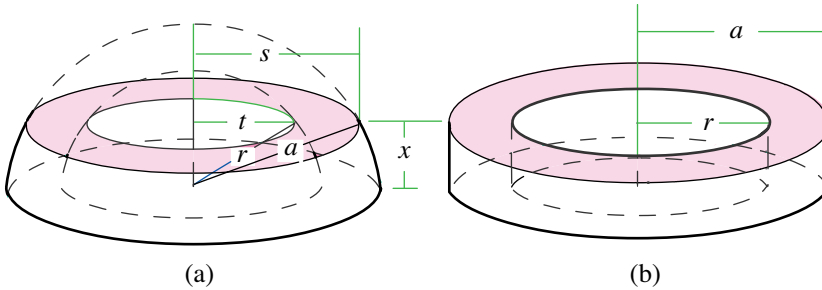


Figure 5.4: The cross-sectional area of a spherical shell cut by a plane that cuts both spheres is constant.

cylindrical shell between two coaxial cylinders with radii r and a, as shown in Figure 5.4b. Therefore, by the slicing principle we have:

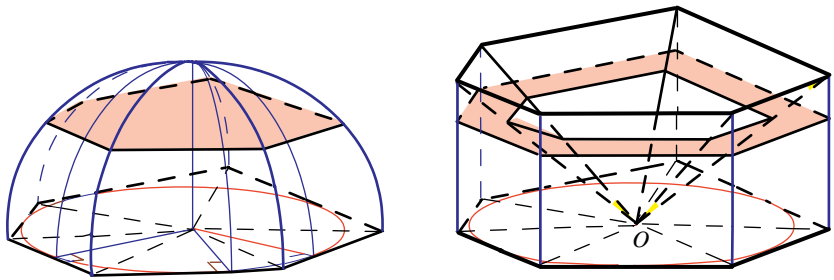
Theorem 5.2. *A slice of a spherical shell between two horizontal planes that cut both spheres has volume equal to the corresponding slice of a cylindrical shell cut by the same planes.*

The volume of the cylindrical shell is the area of its base times its altitude. In terms of the radii, the volume is $\pi(a^2 - r^2)h$, where h is the distance between the parallel cutting planes. Thus, for given radii, the volume is proportional to the distance between the parallel cutting planes.

More generally, the portion of a spherical shell between two horizontal planes has volume equal to the corresponding portion of the punctured cylindrical shell. By considering very thin spherical shells we can use this result to deduce the surface area of a sphere. We prefer to deduce this later in Section 5.6, where we explore surface areas of general Archimedean globes. In the next five sections we do not assume a knowledge of the volume or surface area of a sphere; we realize it could be used to simplify many of our proofs by comparing cross-sectional areas of spheres and circumscribed globes. Our purpose is to present a sequence of elementary results that could have been discovered by Archimedes had he tried to simplify his geometric analysis of wedges.

5.4 VOLUME OF AN ARCHIMEDEAN GLOBE

For convenience in drawing figures, we define first an Archimedean dome circumscribing a hemisphere as in Figure 5.5a. The base is an arbitrary polygon circumscribing the equator. Each cross section of the dome by a plane parallel to the base is a similar polygon circumscribing the circular cross section of the sphere and having the same orientation as the base relative to the polar axis through the center. The union of a dome with its equatorial mirror image creates an Archimedean globe.



(a) Archimedean dome (b) Punctured prismatic container

Figure 5.5: An Archimedean dome whose volume is equal to that of its punctured circumscribing prism.

Incidentally, globes representing the Earth are often made this way using a regular dodecagon as the equatorial base. Astronomical observatories use Archimedean domes to cover their telescopes.

Figure 5.5b shows a circumscribing prism of the dome with the same polygonal base and the same altitude, from which a pyramid with congruent base and vertex O has been removed. The pyramid has volume one-third that of the prism, so the solid that remains has volume two-thirds that of the prism. Each horizontal cross section of the punctured prism is a polygonal ring bounded by two polygons similar to the base. We will show that the area of this ring is equal to the area of the corresponding cross section of the dome, which implies that the dome and the punctured prism have equal volumes.

To show equality of cross-sectional areas, divide the dome into wedges of the type shown in Figure 5.6a, with a right triangular base of altitude a, a circular cylindrical face of radius a, and two vertical plane faces. The curve on the cylindrical face joining the top of the dome to the vertex of the right angle at the base of the triangle is called a *meridian*; it is a quarter of a circle of radius a. The curve on the other edge of the cylindrical face is an ellipse.

The circumscribing prism is correspondingly divided into different triangular prisms of altitude a, each having a base congruent to the corresponding right triangular base of the wedge, as shown in Figure 5.6b. Let T denote the area of a typical triangular base. A horizontal cutting plane at distance x above the base cuts a triangle of area $A(x)$ from the dome and a trapezoid of area $T(x)$ from the punctured prism. It suffices to show that $A(x) = T(x)$.

The area of the trapezoid is equal to T minus the area of a smaller similar triangle of altitude c and similarity ratio c/a, as indicated in Figure 5.6b. But $c/a = x/a$, so the smaller triangle has area $(x/a)^2T$, hence $T(x) = (1 - (x/a)^2)T$.

Let y be the altitude of the triangular cross section of the wedge in Figure 5.6a cut by a plane at a distance x from the base. This triangle is similar to the triangular base with similarity ratio y/a, so its area $A(x)$ is $(y/a)^2T$. But from Figure 5.6a we have $x^2 + y^2 = a^2$ because the meridian is a circular arc, and therefore $(y/a)^2T = (1 - (x/a)^2)T = T(x)$. In other words, $A(x) = T(x)$, as we set

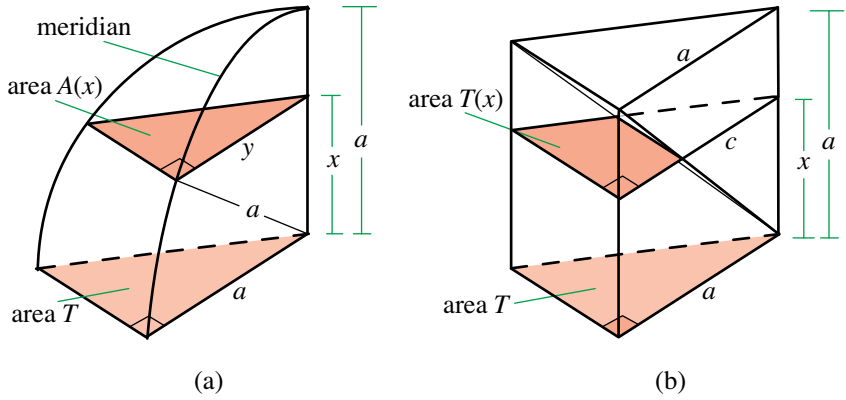


Figure 5.6: Cross-sectional areas $A(x)$ and $T(x)$ are equal for all x.

out to prove. This argument gives us:

Theorem 5.3. (a) *Corresponding slices of the globe and punctured prism cut by two planes parallel to the equator have equal volumes.*
(b) *The volume of an Archimedean globe is two-thirds the volume of its circumscribing prism.*

When the number of edges of the polygonal base tends to ∞, the equatorial polygon becomes a circle, the Archimedean globe becomes a sphere, and the circumscribing prism becomes a circular cylinder. Thus, Theorem 5.1 and Corollary 5.2 are limiting cases of Theorem 5.3.

Examples. Archimedean globes can also be constructed by combining wedge-like portions of n semicircular cylinders whose axes are in the equatorial plane and intersect at the center of the inscribed equator, each axis being parallel to an edge of the polygonal base. The two simplest examples ($n = 3$ and $n = 4$) are shown in

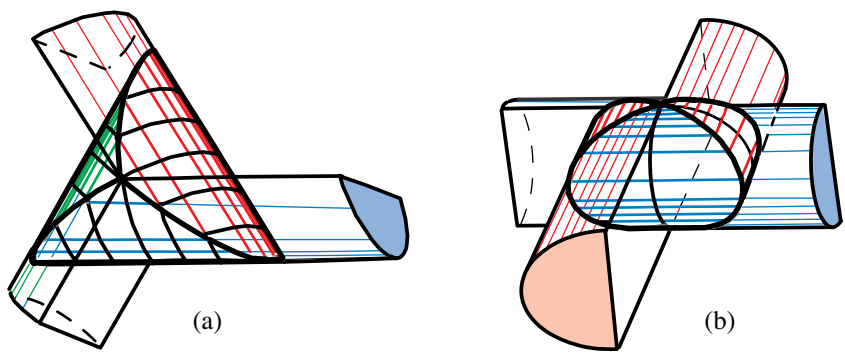


Figure 5.7: Portions of semicircular cylindrical wedges combined to form Archimedean globes.

Figures 5.7a and 5.7b. The solid in Figure 5.7b (usually described as the intersection of two cylinders) has volume two-thirds that of the smallest box that contains it.

We selected the name "Archimedean dome" because of a special case considered by Archimedes. In his preface to *The Method* [47; supplement, p. 12] Archimedes announced (without proof) that the volume of the intersection of two congruent orthogonal circular cylinders is two-thirds the volume of the circumscribing cube. In [47; pp. 48-50], Zeuthen verifies this with the method of centroids and levers employed by Archimedes in treating the sphere. However, if we observe that half the solid of intersection is an Archimedean dome with a square base, and compare its volume with that of its punctured prismatic container, we immediately obtain the two-thirds volume ratio.

As a limiting case, when the polygonal cross sections of an Archimedean dome become circles and the punctured container becomes a circumscribing cylinder punctured by a cone, we obtain a purely geometric derivation of the Archimedes volume ratio for a sphere and cylinder. In fact, an Archimedean globe whose equator is a regular n-gon circumscribing a sphere of radius a has volume $(4/3)na^3\tan(\pi/n)$, whose limiting value as $n \to \infty$ is $4\pi a^3/3$, the volume of a sphere of radius a.

5.5 VOLUME OF AN ARCHIMEDEAN SHELL

Next we analyze shells, which are analogous to solids constructed from wedge-like portions of cylindrical pipes. The volume of the shell between two concentric Archimedean domes is, of course, the difference between the volumes of the outer and inner domes. Theorem 5.3(a) gives parts (a) and (c) of the following theorem:

Theorem 5.4. (a) *Corresponding slices of an Archimedean shell and the punctured circumscribing prism cut by two planes parallel to the equator have equal volumes.*
(b) *A slice of an Archimedean shell between parallel planes that cut both domes has volume equal to the corresponding slice of a prismatic shell (of constant thickness) cut by the same planes. The volume is the product of the distance between the cutting planes and the area of the polygonal ring on the base.*
(c) *Corresponding slices of two Archimedean shells with bases of equal area cut by planes parallel to their common equatorial plane have equal volumes.*

To prove part (b), look at Figure 5.8a, which shows one wedge cut from two concentric Archimedean domes with radii r and a, where $r < a$. The base of the wedge is a trapezoid of altitude $a - r$. The wedge is intersected by two parallel horizontal planes that cut both domes. Horizontal cross sections are trapezoids of variable altitude, each similar to the trapezoidal base.

The corresponding cross sections in Figure 5.8b are trapezoids of equal area but with a fixed altitude $a - r$. Therefore, by the slicing principle, the Archimedean shell and prismatic shell have equal volumes.

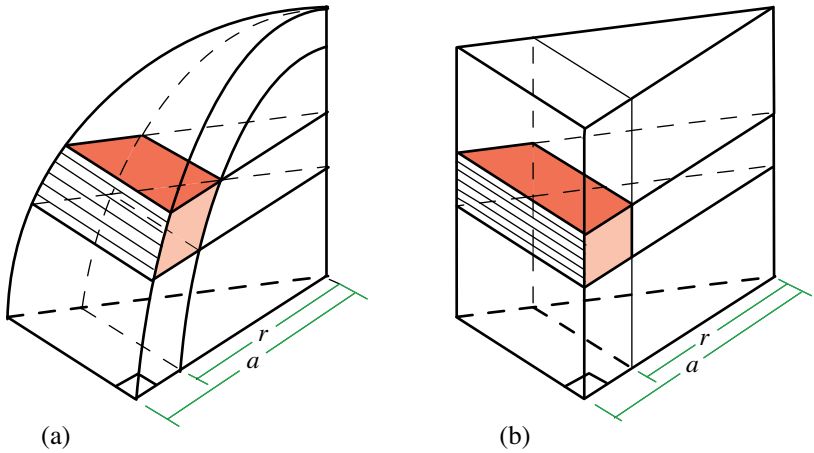


Figure 5.8: Parallel planes cut the Archimedean shell and the prismatic shell in slices of equal volumes.

5.6 SURFACE AREA OF AN ARCHIMEDEAN DOME

Theorem 5.4(b) can be used to give a heuristic argument for determining the surface area of an Archimedean globe. Because of symmetry it suffices to treat the upper dome.

Figure 5.9a shows one wedge of a very thin Archimedean shell, with outer base b, outer radius a, and inner radius r, where r is very nearly equal to a. The shell can be unwrapped (Figure 5.9b) to form a figure that is flat and almost a prism, with its volume equal to the lateral area A of the wedge times its thickness, or $A(a-r)$. We want to determine A.

In proving Theorem 5.4(b) we found that a wedge of an Archimedean shell has volume equal to that of a portion of a prismatic shell of thickness $a-r$. This portion, shown in Figure 5.9c, is nearly a thin rectangular slab of base b, altitude a, thickness $a-r$, and volume $ba(a-r)$. Equating this to $A(a-r)$ we find $A = ba$. The sum of the lateral areas A of all the slices is equal to the sum of the corresponding products ba which, in turn, is the lateral surface area of the circumscribing prism.

The same analysis applies to any portion of the Archimedean shell between two parallel cutting planes. For the limiting case when $r \to a$ we obtain:

Theorem 5.5. (a) *The lateral surface area of a slice of an Archimedean globe between two parallel planes is equal to the lateral surface area of the corresponding slice of the circumscribing prism. The area is proportional to the distance between the parallel cutting planes.*

(b) *The total surface area of an Archimedean globe is equal to the lateral surface area of the circumscribing prism, which is four times the area of the equatorial base.*

This result, discovered by a heuristic argument, can be given a rigorous proof by using the method of exhaustion or integration. In the limiting case when the

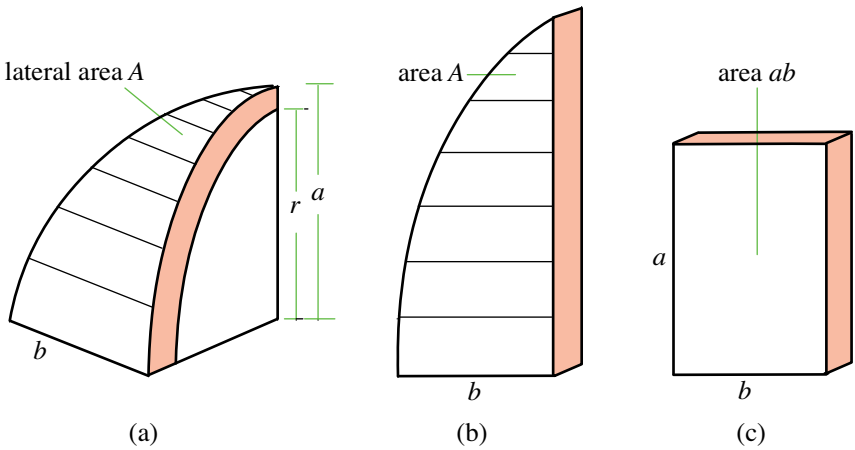


Figure 5.9: The curved face of a slice of a thin Archimedean shell in (a) unwrapped so that it is flat as in (b). The volume of the shell is very nearly equal to the volume of the rectangular slab in (c) of the same thickness.

circumscribing prism becomes a circular cylinder, we obtain:

Corollary 5.2. *The lateral surface area of a spherical slice cut by two parallel planes is equal to the lateral surface area of the corresponding slice of the circumscribing cylinder.*

Using a different approach, Archimedes found the surface area of a sphere [47, p. 39, Proposition 33], and the surface area of a segment of a sphere [47, p. 53, Proposition 43]. The statement for the segment is particularly elegant because it involves only one parameter, the slant height of a cone inscribed in the segment. Proposition 43 states that the surface area of the segment is equal to that of a circle whose radius is the slant height of the cone inscribed in the segment. This result holds more generally for the surface area of a segment of an Archimedean dome (the portion of the surface of the dome above a plane parallel to the equatorial plane, as shown in Figure 5.10a).

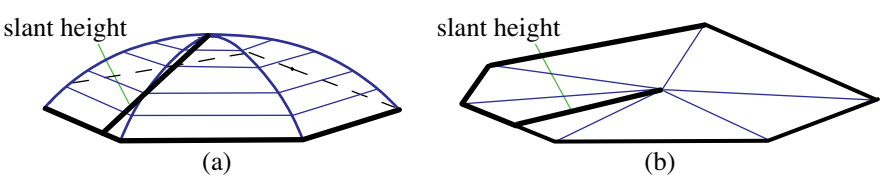


Figure 5.10: The surface area of a segment of a dome is equal to that of a polygon similar to the base.

Theorem 5.6. *The surface area of a segment of an Archimedean dome is equal to that of a polygon similar to the polygonal base circumscribing a circle whose radius is the slant height of the corresponding inscribed pyramid.*

Proof. By Theorem 5.5(a), the surface area of a segment of height h is equal to hp, where p is the perimeter of the polygonal base. Let a be the radius of the equator, and let s denote the slant height of the inscribed pyramid (Figure 5.10a). Then s is the hypotenuse of a right triangle with h as one leg, and s is also one leg of a similar right triangle with hypotenuse $2a$. Therefore $2a/s = s/h$, or $s/a = 2h/s$. But s/a is the similarity ratio of similar polygons circumscribing circles of radii s and a, respectively. The polygon circumscribing the circle of radius s is shown in Figure 5.10b. If p_s denotes its perimeter, then its area is $sp_s/2$. By similarity, $p_s = (s/a)p = (2h/s)p$, so $sp_s/2 = hp$, as required. The relation $s/a = 2h/s$ for the sphere also proves Proposition 43.

Theorem 5.4(b) states that the volume of an Archimedean globe is two-thirds the volume of the smallest circumscribing prism. Now we prove the companion theorem for surface area:

Theorem 5.7. *The surface area of an Archimedean globe is two-thirds the total surface area of its circumscribing prism.*

Proof. By Theorem 5.5(b), the total surface area of an Archimedean globe is equal to the lateral surface area of the smallest circumscribing prism. Therefore to prove Theorem 5.7 it suffices to show that each polygonal base of the prism has area equal to one-fourth the lateral surface area of the prism. Then the areas of the two bases plus the lateral surface area is three-halves the surface area of the inscribed Archimedean globe.

The lateral surface of the prism can be unwrapped to form a rectangle of area $2ap$, where p is the perimeter of the base, and $2a$ is the altitude of the prism. The polygonal base can be divided into right triangles of the type shown in Figure 5.8b, each with altitude a and area $ab_k/2$, where b_k is the base of the triangle. The sum of the b_k is equal to p, and the area of the polygonal base is $ap/2 = (2ap)/4$, as required.

When the polygonal base approaches a circle as a limit we obtain:

Corollary 5.3. (Archimedes) *The surface area of a sphere is two-thirds the total surface area of its circumscribing cylinder, which is four times the area of the equatorial disk.*

Theorems 5.4 and 5.7 provide new proofs and significant generalizations of the landmark discoveries of Archimedes mentioned in the opening sentence of this chapter. As already remarked, Archimedes knew that the volume of intersection of two perpendicular cylinders is two-thirds that of the smallest cube that contains it, but apparently he never considered the corresponding surface areas, which, by Theorems 5.4 and 5.7, are in the same ratio. Finding the volume of two intersecting cylinders has become a standard exercise in calculus texts, but, except for the case of a sphere, we have not seen the corresponding area relation of Theorem 5.7 discussed in the literature.

We turn next to two surprising consequences of Theorem 5.5.

5.7 INCONGRUENT SOLIDS WITH EQUAL VOLUMES AND EQUAL SURFACE AREAS

Figure 5.11a shows a horizontal slice of an Archimedean shell between two parallel planes that cut both the inner and outer domes. Figure 5.11b shows the corresponding prismatic slice of the same constant thickness. The surface of each slice consists of four components: (1) an upper horizontal polygonal ring, (2) a lower horizontal polygonal ring, (3) an outer lateral surface, and (4) an inner lateral surface. We now prove:

Theorem 5.8. *The two slices so described have equal volumes, and corresponding components of the surface of each slice have equal areas. Consequently the two slices have equal total surface areas.*

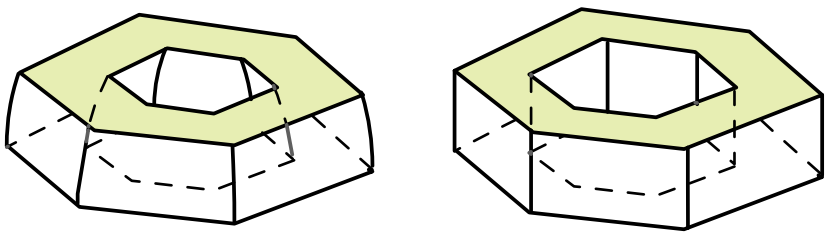


Figure 5.11: Two incongruent solids of equal volume with corresponding area components of equal areas.

Proof. The volumes are equal by Theorem 5.4(b). From the analysis leading to Theorem 5.4, the upper horizontal polygonal rings have equal areas, as do the lower horizontal polygonal rings. By Theorem 5.5(a), the two outer lateral surface areas are equal, as are the two inner lateral surface areas.

Theorem 5.8 provides several infinite families of pairs of incongruent solids that have equal volumes and equal total surface areas. One family is obtained by varying the number of edges or the shape of the equatorial circumscribed polygon, a second by varying the distance between the parallel cutting planes, and a third by varying the distance of one cutting plane from the equatorial plane. Incidentally, the solids in Figure 5.11 resemble washers commonly used, for example, in plumbing fixtures.

5.8 QUADRATURE OF THE SINE CURVE

The next surprising consequence of Theorem 5.5 is the quadrature of the sine curve. A point on a unit circle that subtends an angle of x radians has rectangular coordinates $(\cos x, \sin x)$. Figure 5.12a shows this circle as the base of a right circular cylinder from which a wedge has been cut by a plane through a diameter inclined at an angle of 45° with the base. The point on the cutting plane directly above the point $(\cos x, \sin x)$ on the base has altitude $\sin x$. In Figure 5.12b the lateral surface

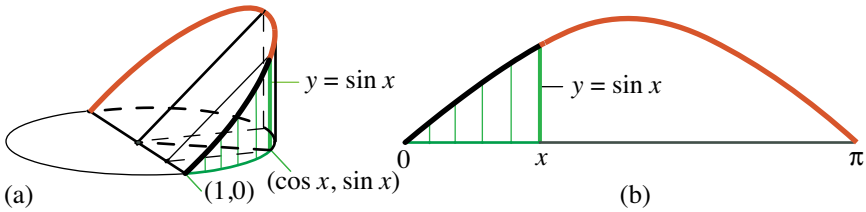


Figure 5.12: Generating a sine curve by cutting a circular cylinder by an inclined plane through a diameter.

of the wedge is unwrapped to form a region lying above an interval of length π (half the circumference of the circle), so the upper boundary of the region has cartesian equation $y = \sin x$. The front half of the wedge in Figure 5.12a can be regarded as a wedge of an Archimedean dome that has been tipped over so that its circular face is in a horizontal plane with the top of the dome at the point with rectangular coordinates $(1, 0)$. The base of the wedge is an isosceles right triangle in a vertical plane. By Theorem 5.5(a) the lateral surface area of any portion of the wedge cut

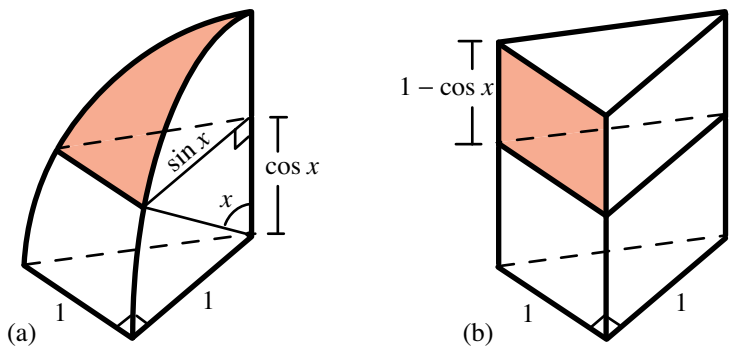


Figure 5.13: The lateral area of a portion of the dome is equal to that of a rectangular face of the smallest circumscribing prism.

by a plane parallel to the base of the dome is equal to the area of the corresponding rectangular face cut from a slice of the smallest circumscribing prismatic shell. If the cutting plane is at a distance $\cos x$ from the base of the dome, as shown in Figure 5.13a, the rectangular face has base 1 and altitude $1 - \cos x$, as shown in Figure 5.13b. Thus, by elementary geometry, we obtain the quadrature of the sine curve if $0 \le x \le \pi/2$, but, of course, the result holds for all real x.

Corollary 5.4. *The area of the region under the sine curve and above the interval* $[0, x]$ *is equal to that of a rectangle of base* 1 *and altitude* $1 - \cos x$.

In calculus notation,

$$\int_0^x \sin t \, dt = 1 - \cos x.$$

5.9 APPLICATION TO CENTROIDS

The method of punctured containers determines the volume of a curved solid in terms of that of a circumscribed punctured prismatic solid whose volume is known or can be easily calculated because it is bounded by plane faces. We cut both solids by horizontal planes that produce cross sections of equal area $A(x)$ at an arbitrary height x above a fixed base. Then we invoke the slicing principle to equate the volumes of the solids cut off between two horizontal planes. In the language of calculus, the value of the integral $\int_{x_1}^{x_2} A(x)\,dx$ is the volume of the portion of each solid cut by all horizontal planes as x varies over some interval $[x_1, x_2]$. (See [1; Theorem 2.7].) Because the integrand $A(x)$ is the same for both solids, the corresponding volumes are equal.

Instead of integrating the common cross sectional area $A(x)$ of two solids to find that their volumes are equal, we could just as well integrate any function $f(x, A(x))$, and the integral over $[x_1, x_2]$ would be the same for both solids. For example, the integral $\int_{x_1}^{x_2} xA(x)\,dx$ is the first moment of the area function $A(x)$ over $[x_1, x_2]$, and this integral divided by the integral $\int_{x_1}^{x_2} A(x)\,dx$ gives the altitude of the centroid of the slice of each solid between the planes $x = x_1$ and $x = x_2$.

Thus, not only are the volumes of the slices equal, but also the altitudes of their centroids are equal. Moreover, for a given k, the moments $\int_{x_1}^{x_2} x^k A(x)\,dx$ with respect to the plane of the base are equal for both slices. In other words, we have:

Theorem 5.9. *With respect to the equatorial plane and a power k, the kth power moments of corresponding slices of an Archimedean shell and its circumscribing punctured prismatic shell are equal.*

We conclude this section with some examples of centroids that can be determined using Theorem 5.9.

Consider a shell between two concentric Archimedean domes with radii r and a, where $r < a$. Theorem 5.9 enables us to locate the centroid of a portion of the shell between two planes parallel to the base that cut both domes. The corresponding slice of a prismatic shell of constant thickness cut by the same planes has its centroid located midway between the two parallel planes. Therefore this is also the height of the centroid of the slice. Hence we have:

Theorem 5.10. *The centroid of a slice of an Archimedean shell between two horizontal planes that cut both domes lies midway between the two planes on the altitude through the common center. In particular, the slice of the shell between the equatorial plane and the plane whose distance from the center is the radius r of the inner dome has its centroid at a distance $r/2$ above the equator.*

In the limiting case when $r \to a$ (so the thickness of the shell tends to 0), we find the following corollary of Theorem 5.10:

Corollary 5.5. *The centroid of the surface of an Archimedean dome is at the midpoint of its altitude.*

In the limiting case when the circumscribing equatorial polygon becomes a circle, this yields a known result for a hemisphere that can be found using surface integrals (see [2; p. 431]). The same limiting case of Theorem 5.10 gives:

Corollary 5.6. *The centroid of the surface of a slice of a sphere* (*in particular, of a segment*)*, is midway between the two parallel cutting planes.*

PART 2: GENERALIZED ARCHIMEDEAN DOMES

The rest of this chapter extends the method of punctured containers by applying it first to general dome-like structures far removed from Archimedean domes, and then to domes with nonuniform mass distributions. The generality of the structures that are amenable to the method of punctured containers is demonstrated by the following examples.

5.10 REDUCIBLE SOLIDS

Cut an Archimedean dome and its punctured container into horizontal slices and assign to each pair of slices the same constant density, which can differ from pair to pair. Because the masses are equal, slice by slice, the total mass of the dome is equal to that of its punctured container, and the centers of mass are at the same height above the base. Or, cut the dome and its punctured container into wedges by vertical half planes through the polar axis, and assign to each pair of wedges the same constant density, which can differ from pair to pair. Again, the masses are equal, wedge by wedge, so the total mass of the dome is equal to that of its punctured container, and the centers of mass are at the same height. Or, imagine an Archimedean dome divided into thin concentric shell-like layers, like those of an onion, each assigned a constant density, which can differ from layer to layer. The punctured container is correspondingly divided into coaxial prismatic layers, each assigned the same constant density as the corresponding shell layer. In this case the masses are equal, shell by shell, so the total mass of the dome is equal to that of its punctured container, and again the centers of mass are at the same height. We are interested in solids with pyramidically punctured prismatic containers that share the following property with Archimedean domes:

Definition (Reducible solid). *A solid is called* reducible *if an arbitrary horizontal slice of the solid and its punctured container have equal volumes, equal masses, and centers of mass at the same height above the base.*

Every uniform Archimedean dome is reducible, and in Section 5.13 we exhibit some nonuniform Archimedean domes that are reducible.

The method of punctured containers enables us to reduce both volume and mass calculations of domes to those of simpler prismatic solids, thus generalizing the profound volume relation between the sphere and cylinder discovered by Archimedes. Another famous result of Archimedes [47; *Method*, Proposition 6] states that the centroid of a uniform solid hemisphere divides its altitude in the ratio 3:5. Using

the method of punctured containers we show that the same ratio holds for uniform Archimedean domes and other more general domes (Theorem 5.17), and we also extend this result to the center of mass of a more general class of nonuniform reducible domes (Theorem 5.18).

In Section 5.11 we dilate an Archimedean dome in a vertical direction to produce a dome with elliptic profiles, then we replace its base by an arbitrary polygon, not necessarily convex. This leads naturally to domes with arbitrary curved bases. Such domes and their punctured prismatic containers have equal volumes and equal moments relative to the plane of the base because of the slicing principle, but if the domes do not circumscribe hemispheres the corresponding lateral surface areas will not be equal. The remainder of this chapter relaxes the requirement of equal surface areas and concentrates on solids having the same volume and moments as their punctured prismatic containers. Section 5.11 treats reducible domes and shells with polygonal bases, then Section 5.12 extends the results to domes with curved bases, and formulates reducibility in terms of mappings that preserve volumes and moments.

The full power of the method of punctured containers is revealed by the treatment of nonuniform mass distributions in Section 5.13. Problems of calculating masses and centroids of nonuniform wedges, shells, and their slices with elliptic profiles, including those with cavities, are reduced to those of simpler punctured prismatic containers. Section 5.14 gives explicit formulas for volumes and centroids, and Section 5.15 reveals the surprising fact that uniform domes with elliptic profiles are essentially the only solids that are reducible to their punctured containers.

5.11 POLYGONAL ELLIPTIC DOMES AND SHELLS

To easily construct a more general class of reducible solids, start with any Archimedean dome and dilate it and its punctured container in a vertical direction by the same scaling factor $\lambda > 0$. The circular cylindrical wedges in Figure 5.5a become elliptic cylindrical wedges, as typified by the example in Figure 5.14a. A circular arc of radius a is dilated into an elliptic arc with horizontal semiaxis a and vertical semiaxis λa. Dilation changes the altitude of the prismatic wedge from a to λa (Figure 5.14b). The punctured container is again a prism punctured by a pyramid. Each horizontal plane at a given height above the base cuts both the elliptic wedge and the corresponding punctured prismatic wedge in cross sections of equal area. Consequently, any two horizontal planes cut both solids in slices of equal volume.

If the elliptic and prismatic wedges have the same constant density, then they also have the same mass, and their centers of mass are at the same height above the base. Thus we have:

Theorem 5.11. *Every uniform elliptic cylindrical wedge is reducible.*

Now assemble a finite collection of nonoverlapping elliptic cylindrical wedges with their horizontal semi axes, possibly of different lengths, in the same horizontal plane, but with a common vertical semiaxis, which meets the base at a common point O called the center. We assume that the density of each component wedge

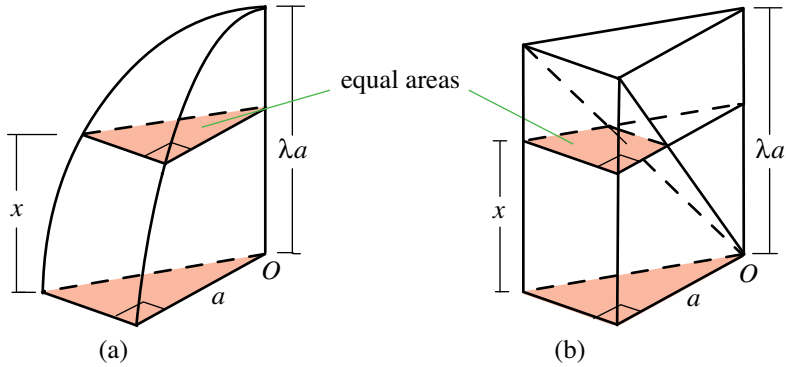


Figure 5.14: (a) Vertical dilation of a cylindrical wedge by a factor λ. (b) Its punctured prismatic container.

is constant, although this constant may differ from component to component. For each wedge, the punctured circumscribing prismatic container with the same density is called its *prismatic counterpart*. The punctured containers assembled in the same manner produce the counterpart of the wedge assemblage. We call an assemblage *nonuniform* if some of its components can have different constant densities. This includes the special case of a uniform assemblage where all components have the same constant density. Because each wedge is reducible we get:

Corollary 5.7. *A nonuniform assemblage of elliptic cylindrical wedges is reducible.*

Polygonal elliptic domes.

Because the base of a finite assemblage is a polygon (a union of triangles with a common vertex O) we call the assemblage *a polygonal elliptic dome*. The polygonal base need not circumscribe a circle and it need not be convex. Corollary 5.7 gives us:

Corollary 5.8. *The volume of a polygonal elliptic dome is equal to the volume of its circumscribing punctured prismatic container, that is, two-thirds the volume of the unpunctured prismatic container, which is the area of the base times the height.*

In the special limiting case when the equatorial polygonal base of the dome turns into an ellipse with center at O, the dome becomes half an ellipsoid, and the circumscribing prism becomes an elliptic cylinder. In this limiting case, Corollary 5.8 reduces to:

Corollary 5.9. *The volume of an ellipsoid is two-thirds that of its circumscribing elliptic cylinder.*

In particular, we get Archimedes' spheroid result [44; *Method*, Proposition 3]:

Corollary 5.10. (Archimedes) *The volume of an ellipsoid of revolution is two-thirds that of its circumscribing circular cylinder.*

Polygonal elliptic shells.

A polygonal elliptic shell is the solid between two concentric similar polygonal elliptic domes. From Theorem 5.11 we obtain:

Theorem 5.12. *The following solids are reducible:*
(a) *A uniform polygonal elliptic shell.*
(b) *A wedge of a uniform polygonal elliptic shell.*
(c) *A horizontal slice of a wedge of type* (b).
(d) *A nonuniform assemblage of shells of type* (a).
(e) *A nonuniform assemblage of wedges of type* (b).
(f) *A nonuniform assemblage of slices of type* (c).

By using as building blocks horizontal slices of wedges cut from a polygonal elliptic shell, we can see intuitively how one might construct, from such building blocks, very general polygonal elliptic domes that are reducible and have more or less arbitrary mass distribution. By considering limiting cases of polygonal bases with many edges, and building blocks with very small side lengths, we can imagine elliptic shells and domes whose bases are more or less arbitrary plane regions, for example, elliptic, parabolic, or hyperbolic segments. The next section describes an explicit construction of general reducible domes with curvilinear bases.

5.12 GENERAL ELLIPTIC DOMES

Replace the polygonal base by a plane region bounded by a curve whose polar coordinates (r, θ) relative to a center O satisfy an equation $r = \rho(\theta)$, where ρ is a piecewise continuous function, and θ varies over an interval of length 2π. Above this base we build an elliptic dome as follows. First, the altitude of the dome is a segment of height $h > 0$ along the polar axis perpendicular to the base at O. We assume that each vertical half plane through the polar axis at angle θ cuts the surface of the dome along a quarter of an ellipse with horizontal semiaxis $\rho(\theta)$ and the same vertical semiaxis h, as in Figure 5.15a.

The ellipse will be degenerate at points where $\rho(\theta) = 0$. Thus, an elliptic wedge is a special case of an elliptic dome. When $\rho(\theta) > 0$, the cylindrical coordinates (r, θ, z) of points on the surface of the dome satisfy the equation of an ellipse:

$$\left(\frac{r}{\rho(\theta)}\right)^2 + \left(\frac{z}{h}\right)^2 = 1. \tag{5.1}$$

The dome is circumscribed by a cylindrical solid of altitude h whose base is congruent to that of the dome (Figure 5.15b). Incidentally, we use the term "cylindrical solid" with the understanding that the solid is a prism when the base is polygonal.

Each point (r', θ', z') on the lateral surface of the cylinder in Figure 5.15b is related to the corresponding point (r, θ, z) on the surface of the dome by

$$\theta' = \theta, \;\; z' = z, \;\; r' = \rho(\theta).$$

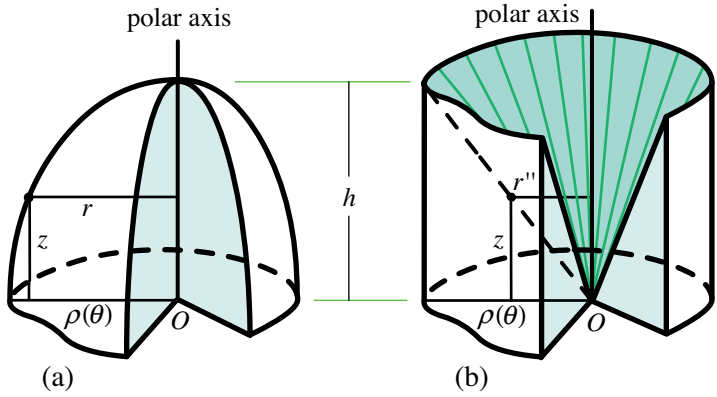


Figure 5.15: An elliptic dome (a), and its circumscribing punctured prismatic container (b).

From this cylindrical solid we remove a conical solid whose surface points have cylindrical coordinates (r'', θ, z), where $z/h = r''/\rho(\theta)$, or

$$r'' = z\rho(\theta)/h.$$

When $z = h$, this becomes $r'' = \rho(\theta)$, so the base of the cone is congruent to the base of the elliptic dome. When the base is polygonal, the conical solid is a pyramid.

More reducible domes.

The polar axis of an elliptic dome depends on the location of center O. For a curvilinear base, we can move O to any point inside the base, or even to the boundary. Moving O will change the function $\rho(\theta)$ describing the boundary of the base, with a corresponding change in the shape of the ellipse determined by (5.1). Thus, the construction generates not one, but infinitely many elliptic domes with a given base.

For any such dome, we can generate another family as follows: Imagine the dome and its prismatic counterpart made up of very thin horizontal layers, like two stacks of cards. Deform each solid by a horizontal translation and rotation of each horizontal layer. The shapes of the solids will change, but their cross-sectional areas will not change. In general, such a deformation may alter the shape of each ellipse on the surface to some other curve, and the deformed dome will no longer be elliptic. The same deformation applied to the prismatic counterpart will change the punctured container to a nonprismatic punctured counterpart. Nevertheless, all the results of this chapter (with the exception of Theorem 5.21) will hold for such deformed solids and their counterparts.

However, if the deformation is a linear shearing that leaves the base fixed but translates each layer by a distance proportional to its distance from the base, then

straight lines are mapped onto straight lines and the punctured prismatic solid is deformed into another prism punctured by a pyramid with the same base. The correspondingly sheared dome will be elliptic because each elliptic curve on the surface of the dome is deformed into an elliptic curve. To visualize a physical model of such a shearing, imagine a general elliptic dome and its counterpart sliced horizontally to form stacks of cards. Pierce each stack by a long pin along the polar axis, and let O be the point where the tip of the pin touches the base. Tilting the pin away from the vertical polar axis, keeping O fixed, results in horizontal linear shearing of the stacks and produces infinitely many elliptic domes, all with the same polygonal base. The prismatic containers are correspondingly tilted, and the domes are reducible.

Reducibility mapping.

For an elliptic dome, we call the corresponding circumscribing punctured cylindrical solid its *punctured container*. Our goal is to show that every uniform elliptic dome is reducible. This will be deduced from a more profound property, stated below in Theorem 5.13. It concerns a mapping that relates elliptic domes and their punctured containers.

To determine the mapping, regard the dome as a collection of layers of similar elliptic domes, like layers of an onion. Choose O as the center of similarity, and for each scaling factor $\mu \leq 1$, imagine a surface $E(\mu)$ such that a vertical half plane through the polar axis at angle θ intersects $E(\mu)$ along a quarter of an ellipse with semiaxes $\mu\rho(\theta)$ and μh. When $\rho(\theta) > 0$, the coordinates r and z of points on this similar ellipse satisfy

$$\left(\frac{r}{\mu\rho(\theta)}\right)^2 + \left(\frac{z}{\mu h}\right)^2 = 1. \tag{5.2}$$

Regard the punctured container as a collection of coaxial layers of similar punctured cylindrical surfaces $C(\mu)$. It is easy to relate the cylindrical coordinates (r', θ', z') of each point on $C(\mu)$ to the coordinates (r, θ, z) of the corresponding point on $E(\mu)$. First, we have

$$\theta' = \theta, \;\; z' = z, \;\; r' = \mu\rho(\theta). \tag{5.3}$$

From (5.2) we find $r^2 + z^2\rho(\theta)^2/h^2 = \mu^2\rho(\theta)^2$, so (5.3) becomes

$$\theta' = \theta, \;\; z' = z, \;\; r' = \sqrt{r^2 + z^2\rho(\theta)^2/h^2}. \tag{5.4}$$

The three equations in (5.4), which are independent of μ, describe a mapping from each point (r, θ, z) of the solid elliptic dome, not on the polar axis, to the corresponding point (r', θ', z') on its punctured container. On the polar axis, $r = 0$ and θ is undefined.

Using (5.2) in (5.4) we obtain (5.3), so points on the ellipse described by (5.2) are mapped onto the vertical segment of length μh through the base point $(\mu\rho(\theta), \theta)$.

Preservation of volumes.

Now we show that the mapping (5.4) preserves volumes. In the $(r,\, \theta,\, z)$ system the volume element is $r\, dr\, d\theta\, dz$, while that in the $(r',\, \theta',\, z')$ system is $r' dr' d\theta' dz'$. From (5.4) we have

$$(r')^2 \;=\; r^2 \;+\; z^2 \rho\,(\theta)^2\,/h^2$$

which, for fixed z and θ, gives $r'dr' \;=\; rdr$. From (5.4) we also have $d\theta' = d\theta$ and $dz' = dz$, so the volume elements are equal:

$$r\, dr\, d\theta\, dz = r' dr' d\theta' dz'.$$

This proves:

Theorem 5.13. *The mapping* (5.4)*, from a general elliptic dome to its punctured prismatic container, preserves volumes. In particular, every uniform elliptic dome is reducible.*

As an immediate consequence of Theorem 5.13 we obtain:

Corollary 5.11. *The volume of a general elliptic dome is equal to the volume of its circumscribing punctured cylindrical container, that is, two-thirds the volume of the circumscribing unpunctured cylindrical container which is simply the area of the base times the height.*

The same formulas show that for a fixed altitude z, we have $r\, dr\, d\theta = r' dr' d\theta'$. In other words, the mapping also preserves areas of horizontal cross sections cut from the elliptic dome and its punctured container. This also implies Corollary 5.11 because of the slicing principle.

Lambert's classical mapping as a special case.

Our mapping (5.4) generalizes Lambert's classical mapping [31], which is effected by wrapping a tangent cylinder about the equator, and then projecting the surface of the sphere onto the cylinder by rays through the axis that are parallel to the equatorial plane. Lambert's mapping takes points on the spherical surface (not at the north or south pole) and maps them onto points on the lateral cylindrical surface in a way that preserves areas. The mapping (5.4) takes each point of a solid sphere (except the polar axis) and maps it onto a point of the punctured solid cylinder in a way that preserves volumes. Moreover, analysis of a thin shell (similar to that in Section 5.6) shows that our mapping also preserves areas when the surface of an Archimedean dome is mapped onto the lateral surface of its prismatic container. Consequently, we have:

Theorem 5.14. *The mapping* (5.4)*, from the surface of an Archimedean dome onto the lateral surface of its prismatic container, preserves areas.*

In the limiting case when the Archimedean dome becomes a hemisphere we get:

Corollary 5.12. (Lambert) *The mapping* (5.4)*, from the surface of a sphere to the lateral surface of its tangent cylinder, preserves areas.*

If the hemisphere in the limiting case has radius a, it is easily verified that (5.4) reduces to Lambert's mapping: $\theta' = \theta,\;\; z' = z,\;\; r' = a$.

5.13 NONUNIFORM ELLIPTIC DOMES

The mapping (5.4) takes each point P of an elliptic dome and carries it onto a point P' of its punctured container. Imagine an arbitrary mass density assigned to P, and assign the same mass density to its image P'. If a set of points P fills out a portion of the dome of volume v and total mass m, say, then the image points P' fill out a solid, which we call the *counterpart*, having the same volume v and the same total mass m. This can be stated as an extension of Theorem 5.13:

Theorem 5.15. *Any portion of a general nonuniform elliptic dome is reducible.*

By analogy with Theorem 5.13, we can say that the mapping (5.4) with weights also preserves masses.

Next we describe specific assignments of variable mass density.

Elliptic shells and elliptic fibers.

A simple way of doing this is to divide a general elliptic dome into thin concentric shell-like layers, like those of an onion, each assigned a constant density, which can differ from layer to layer. The punctured container is correspondingly divided into coaxial prismatic layers, each assigned the same constant density as the corresponding shell layer. In this case the masses are equal shell by shell, the total mass of the dome is equal to that of its punctured container, and again the centers of mass are at the same height.

More precisely, we exploit the structure of the dome as a collection of similar domes, as follows.

Take a general elliptic dome of altitude h, and denote its elliptic surface by $E(1)$. Scale $E(1)$ by a positive factor $\mu < 1$ to produce a similar elliptic surface $E(\mu)$. The region between two surfaces $E(\mu)$ and $E(\nu)$ is called an *elliptic shell*.

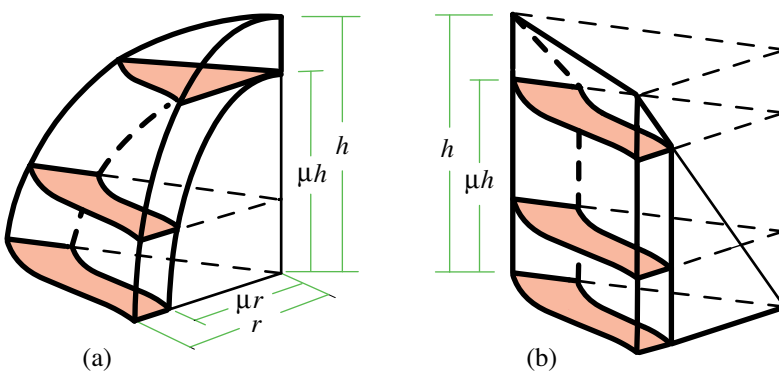


Figure 5.16: Elliptic shell (a), and its counterpart (b).

Each base in the equatorial plane is bounded by portions of two curves with polar equations $r = \mu\rho(\theta)$ and $r = \nu\rho(\theta)$, and two segments with $\theta = \theta_1$ and

$\theta = \theta_2 > \theta_1$. If $\theta_2 - \theta_1 < \pi$ the shell has the shape of a wedge subtending an angle $\theta_2 - \theta_1$. A wedge with a small angle is said to be *sharp*.

Figure 5.16a shows an example of an elliptic shell between $E(1)$ and $E(\mu)$, and Figure 5.16b shows its counterpart. Each base in the equatorial plane is bounded by portions of two curves with polar equations $r = \rho(\theta)$ and $r = \mu\rho(\theta)$, and two segments with $\theta = \theta_1$ and $\theta = \theta_2$.

An elliptic shell between $E(1)$ and $E(\mu)$ is regarded as an elliptic dome with a cavity, or, equivalently, as an elliptic dome with density 0 everywhere inside $E(\mu)$.

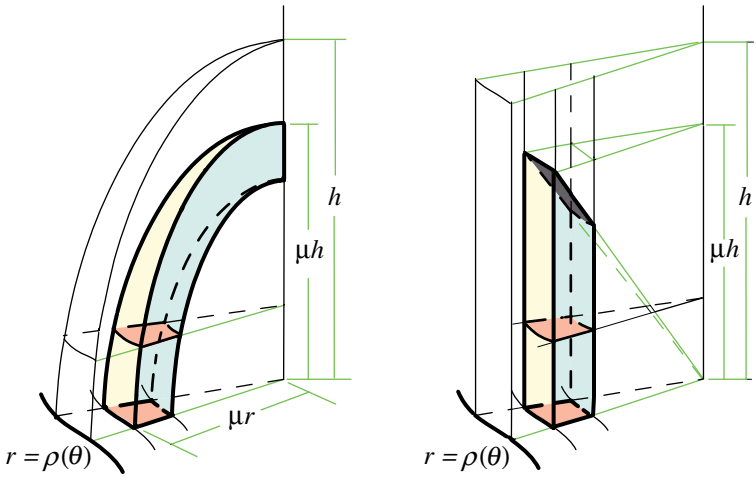


Figure 5.17: Elliptic fiber element (a) and its counterpart (b).

The shaded region in Figure 5.17a is a sharp wedge of an elliptic shell. The outer curved face lies on surface $E(\mu)$ and the inner face lies on another surface, say $E(\nu)$, where $0 < \nu < \mu$. If ν is nearly equal to μ, we call the region an *elliptic fiber*. Figure 5.17b shows the counterpart of this fiber, a vertical prismatic solid. If the fiber is uniform, so is its counterpart, with the same constant density; they have the same mass, and their centers of mass are at the same height above the base.

Elliptic fibers can be used as building blocks to construct many types of nonuniform elliptic domes with variable density. We simply assign to each fiber a constant density, which can differ from fiber to fiber. For example, assemblages of fibers can produce elliptic shells of variable density inherited from the fibers. These can be assembled to form nonuniform elliptic domes. By using fibers with arbitrarily small bases, we can assign arbitrary mass density to the points of a general elliptic dome and its punctured container so that any portion of the dome is reducible. In particular, the portion and its counterpart have equal mass.

In this way, by a limiting process, we produce a reducible nonuniform elliptic dome. We call such a dome *fiber-elliptic*.

An equivalent procedure is to assign an arbitrary mass density $f(r, \theta)$ to the base of a dome. Consider a thin fiber that emanates from a point $(\mu\rho(\theta), \theta)$ on the

base, and assign density $f(\mu\rho(\theta, \theta)$ to each point of the fiber. In other words, the mass density along the elliptic fiber has a constant value inherited from the point at which the fiber meets the base. Of course, the constant may differ from point to point on the base. The elliptic fiber maps into a vertical fiber in the punctured container (of length μh, where h is the altitude of the dome), with the same mass density $f(\mu\rho(\theta), \theta)$.

An important special case occurs when the assigned density is constant along the base curve $r = \rho(\theta)$ and along each curve $r = \mu\rho(\theta)$ similar to the base curve, the density depending only on μ. Then each surface $E(\mu)$ will have constant density. We call domes with this assignment of mass density *shell-elliptic*. For fiber-elliptic and shell-elliptic domes, horizontal slices cut from any portion of the dome and its counterpart have equal masses, and their centers of mass are at the same height above the base, as a consequence of Theorem 5.15. Thus, we have:

Corollary 5.13. *Any portion of a fiber-elliptic dome is reducible. Any portion of a shell-elliptic dome is reducible. In particular, a sphere with spherically symmetric mass distribution is reducible.*

In the same way, if we build a nonuniform shell-elliptic dome with a finite number of similar elliptic shells, each with its density inherited from the base, then any horizontal slice pierced by the cavity has its center of mass midway between the two horizontal cutting planes. Moreover, the following theorem holds for every such shell-elliptic dome.

Theorem 5.16. *A horizontal slice pierced by the cavity of a nonuniform shell-elliptic dome has volume and mass equal, respectively, to those of its prismatic counterpart. Each volume and mass is independent of the height above the base and each is proportional to the thickness of the slice. Consequently, the center of mass of such a slice lies midway between the two cutting planes.*

Corollary 5.14. (Sphere with cavity) *For a spherically symmetric distribution of mass inside a sphere with a concentric cavity, any slice between parallel planes pierced by the cavity has volume and mass proportional to the thickness of the slice, and is independent of the location of the slice.*

Corollary 5.14 implies that the one-dimensional vertical projection of the density is constant along the cavity. This simple result has profound consequences in tomography, which deals with the inverse problem of reconstructing spatial density distributions from a knowledge of their lower-dimensional projections. Details of this application will appear later, in Chapter 15.

The reducibility properties of an elliptic dome also hold for the more general case in which we multiply the mass density $f(\mu\rho(\theta), \theta)$ by any function of z. Such change of density could be imposed, for example, by an external field (such as atmospheric density in a gravitational field that depends only on height z). Consequently, not only are the volume and mass of any portion of this type of nonuniform elliptical dome equal to those of its counterpart, but the same is true for all moments with respect to the horizontal base.

5.14 FORMULAS FOR VOLUME AND CENTROID

This section uses reducibility to give formulas for volumes and centroids of various building blocks of elliptic domes with an arbitrary curvilinear base.

Volume of an elliptic shell.

We begin with the simplest case. Cut a piece (call it a wedge) of an elliptic dome of altitude h by two vertical planes through the polar axis, then remove a similar wedge scaled by a factor μ, where $0 < \mu < 1$, as shown in Figure 5.17a. Assume the unpunctured cylindrical container in Figure 5.17b has volume V. By Corollary 5.11 the outer wedge has volume $2V/3$, and the similar inner wedge has volume $2\mu^3 V/3$, so the volume v of the shell element and its prismatic counterpart is the difference

$$v = \frac{2}{3}V(1-\mu^3). \tag{5.5}$$

Now, $V = Ah$, where A is the area of the base of both the elliptic wedge and its container. The base of the elliptic shell and its unpunctured container have area $B = A - \mu^2 A$, so $A = B/(1-\mu^2)$, $V = Bh/(1-\mu^2)$, and (5.5) can be written as

$$v = \frac{2}{3}Bh\frac{1-\mu^3}{1-\mu^2} = \frac{2}{3}Bh(1+\frac{\mu^2}{1+\mu}). \tag{5.6}$$

This also holds for the total volume of an assemblage of elliptic shells with a given h and μ, with B representing the total base area. The product Bh is the volume of the corresponding unpunctured cylindrical container of altitude h, so (5.6) gives us

$$v_\mu(h) = \frac{2}{3}v_{cyl}(1+\frac{\mu^2}{1+\mu}), \tag{5.7}$$

where $v_\mu(h)$ is the volume of the assemblage of elliptic shell elements and of the counterpart, and v_{cyl} is the volume of its unpunctured cylindrical container.

When $\mu = 0$ in (5.7), the assemblage of elliptic wedges has volume $v_0(h) = 2v_{cyl}/3$, so we can write (5.7) in the form

$$v_\mu(h) = v_0(h)(1+\frac{\mu^2}{1+\mu}), \tag{5.8}$$

where $v_0(h)$ is the volume of the outer dome of the assemblage and its counterpart. If μ approaches 1 the shell becomes very thin, the ratio $\mu^2/(1+\mu)$ approaches $1/2$, and (5.7) shows that $v_\mu(h)$ approaches v_{cyl}. In other words, a very thin elliptic shell element has volume very nearly equal to that of its very thin unpunctured cylindrical container. An Archimedean shell has constant thickness equal to that of the prismatic container, so the lateral surface area of an assemblage of Archimedean wedges is equal to the lateral surface area of its prismatic container (see Theorem 5.5b). This argument does not apply to a nonspherical elliptic shell because it does not have constant thickness.

Next we derive a formula for the height of the centroid of a uniform elliptic wedge above the plane of its base.

Theorem 5.17. *A uniform elliptic wedge or dome of altitude h has volume two-thirds that of its unpunctured prismatic container. Its centroid is located at height c above the plane of the base, where*

$$c = \frac{3}{8}h. \tag{5.9}$$

Proof. It suffices to prove (5.9) for the prismatic counterpart. For a prism of altitude h, the centroid is at a distance $h/2$ above the plane of the base. For a cone or pyramid with the same base and altitude it is known that the centroid is at a distance $3h/4$ from the vertex. To determine the height c of the centroid of a punctured prismatic container above the plane of the base, assume the unpunctured prismatic container has volume V and equate moments to get

$$c(\frac{2}{3}V) + \frac{3h}{4}(\frac{1}{3}V) = \frac{h}{2}V,$$

from which we find (5.9). By Theorem 5.15, the centroid of the inscribed elliptic wedge is also at a height $3h/8$ above the base. The result is also true for any uniform elliptic dome formed as an assemblage of wedges.

Equation (5.9) is equivalent to saying, in the style of Archimedes, that the centroid divides the altitude in the ratio 3:5.

Corollary 5.15. (a) *The centroid of a uniform Archimedean dome divides its altitude in the ratio* 3:5.

(b) (Archimedes) *The centroid of a uniform hemisphere divides its altitude in the ratio* 3:5.

Formula (5.9) is obviously true for the center of mass of a nonuniform assemblage of elliptic wedges of altitude h, each with its own constant density.

Centroid of an elliptic shell.

Now we can find, for an elliptic shell between $E(1)$ and $E(\mu)$, the height $c_\mu(h)$ of its centroid above the plane of its base. The volume and centroid results are summarized as:

Theorem 5.18. *A nonuniform assemblage of elliptic shells with common altitude h and scaling factor μ has volume $v_\mu(h)$ given by* (5.8). *The height $c_\mu(h)$ of the centroid above the plane of its base is given by*

$$c_\mu(h) = \frac{3}{8}h(1 + \frac{\mu^3}{1+\mu+\mu^2}). \tag{5.10}$$

Proof. Consider first a single uniform elliptic shell element. It suffices to do the calculation for the prismatic counterpart. The inner wedge has altitude μh, so by (5.9) its centroid is at height $3\mu h/8$. The centroid of the outer wedge is at height $3h/8$. If the outer wedge has volume V_{outer}, the inner wedge has volume $\mu^3 V_{\text{outer}}$, and the shell element between them has volume $(1-\mu^3)V_{\text{outer}}$. Equating moments and canceling the common factor V_{outer} we find

$$(\frac{3}{8}\mu h)\mu^3 + c_\mu(h)(1-\mu^3) = \frac{3}{8}h,$$

from which we obtain (5.10), which also holds for any nonuniform assemblage of elliptic shell elements with the same h and μ, each of constant density, although the density can differ from element to element.

When $\mu = 0$, (5.10) gives $c_0(h) = 3h/8$.

When $\mu \to 1$, the shell becomes very thin and the limiting value of $c_\mu(h)$ in (5.10) is $h/2$. This also follows from Theorem 5.16 when the shell is very thin and the slice includes the entire dome. It is also consistent with Corollary 5.5, which states that the centroid of the surface of an Archimedean dome is at the midpoint of its altitude.

We leave it as an exercise for the reader to show that the height $c_{\mu\nu}(h)$ of the centroid of a shell between $E(\mu)$ and $E(\nu)$ is equal to

$$c_{\mu\nu}(h) = \frac{3}{8}h(\nu + \frac{\mu^3}{\nu^2+\nu\mu+\mu^2}).$$

Centroid of a slice of a uniform elliptic dome.

More generally, we can determine the centroid of a slice of altitude z of a uniform elliptic wedge. By reducing this calculation to that of the prismatic counterpart, shown in Figure 5.18, the analysis becomes simple. For clarity, the base in Figure 5.18 is shown as a triangle, but the same argument applies to a more general base

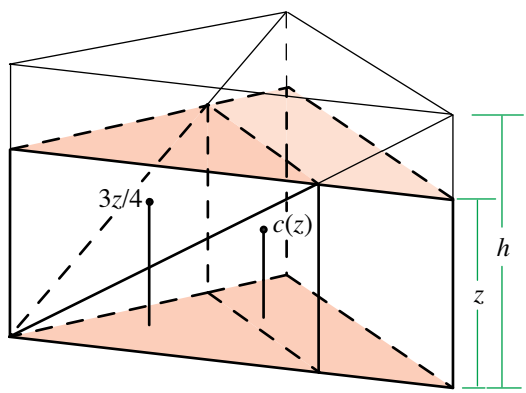


Figure 5.18: Calculating the centroid of a slice of altitude z cut from a wedge of altitude h.

like that in Figure 5.17. The slice in question is obtained from a prism of altitude z and volume $V(z) = \lambda V$, where V is the volume of the unpunctured prismatic container of altitude h, and $\lambda = z/h$. The centroid of the slice is at an altitude $z/2$ above the base. We remove from this slice a pyramidal portion of altitude z and

volume $v(z) = \lambda^3 V/3$, whose centroid is at an altitude $3z/4$ above the base. The portion that remains has volume

$$V(z) - v(z) = (\lambda - \frac{1}{3}\lambda^3)V \tag{5.11}$$

and centroid at altitude $c(z)$ above the base.

To determine $c(z)$, equate moments to obtain

$$\frac{3z}{4}v(z) + c(z)(V(z) - v(z)) = \frac{z}{2}V(z),$$

which gives

$$c(z) = \frac{\frac{z}{2}\mathrm{V}(z) - \frac{3z}{4}v(z)}{V(z) - v(z)}.$$

Because of the relations $V(z) = \lambda^3 V/3$, and $v(z) = \lambda^3 V/3$, we obtain:

Theorem 5.19. *A slice of altitude z cut from a uniform elliptic wedge of altitude h has volume given by* (5.11)*, where $\lambda = z/h$ and V is the volume of the unpunctured prismatic container. The height $c(z)$ of the centroid is given by*

$$c(z) = \frac{3}{4}z\frac{2 - \lambda^2}{3 - \lambda^2}. \tag{5.12}$$

When $z = h$ then $\lambda = 1$ and this reduces to (5.9). For small z the right member of (5.12) is asymptotic to $z/2$. This is reasonable because for small z the walls of the dome are nearly perpendicular to the plane of the equatorial base, so the dome is almost cylindrical near the base.

Centroid of a slice of a shell.

There is a common generalization of (5.10) and (5.12). Cut a slice of altitude z from a shell having altitude h and scaling factor μ, and let $c_\mu(z)$ denote the height of its centroid above the base. Again, we simplify the calculation of $c_\mu(z)$ by reducing it to that of its prismatic counterpart. The slice in question is obtained from an unpunctured prism of altitude z whose centroid has altitude $z/2$ above the base. As in Theorem 5.19, let $\lambda = z/h$. If $\lambda \le \mu$, the slice lies within the cavity, and the prismatic counterpart is the same unpunctured prism of altitude z, in which case we know from Theorem 5.16 that

$$c_\mu(z) = \frac{z}{2} \quad (\lambda \le \mu). \tag{5.13}$$

If $\lambda \ge \mu$, the slice cuts the outer elliptic dome as shown in Figure 5.19a. In this case the counterpart slice has a slant face due to a piece removed by the puncturing pyramid, as indicated in Figure 5.19b.

Let V denote the volume of the unpunctured prismatic container of the outer dome. Then λV is the volume of the unpunctured prism of altitude z. Remove from

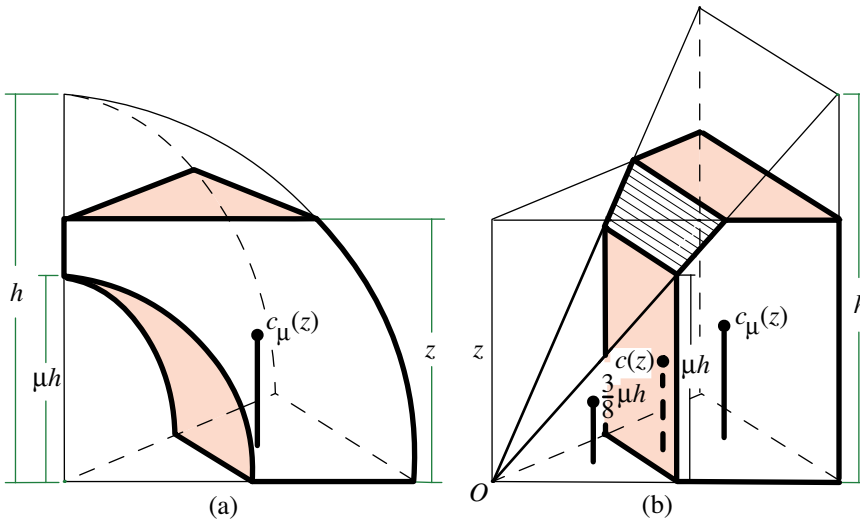


Figure 5.19: Determining the centroid of a slice of altitude $z \geq \mu h$ cut from an elliptic shell element.

this prism the puncturing pyramid of volume $\lambda^3 V/3$, leaving a solid whose volume is

$$V(z) = \lambda V - \frac{1}{3}\lambda^3 V \qquad (\lambda \geq \mu) \tag{5.14}$$

and whose centroid is at the altitude $c(z)$ given by (5.12). This solid is the union of the counterpart slice in question, and an adjacent pyramid with vertex O, altitude μh, volume

$$v_\mu = \frac{2}{3}\mu^3 V, \tag{5.15}$$

and centroid at altitude $3\mu h/8$. The counterpart slice has volume

$$V(z) - v_\mu = (\lambda - \frac{1}{3}\lambda^3 - \frac{2}{3}\mu^3)V. \tag{5.16}$$

To find the altitude $c_\mu(z)$ of its centroid, equate moments to obtain

$$(\frac{3}{8}\mu h)v_\mu + c_\mu(z)(V(z) - v_\mu) = c(z)V(z),$$

from which we find

$$c_\mu(z) = \frac{c(z)V(z) - (\frac{3}{8}\mu h)v_\mu}{V(z) - v_\mu}.$$

Now we use (5.12), (5.14), (5.15) and (5.16). After some simplification we find the result

$$c_\mu(z) = \frac{3}{4}h\,\frac{\lambda^2(2-\lambda^2) - \mu^4}{\lambda(3-\lambda^2) - 2\mu^3} \qquad (\lambda \geq \mu). \tag{5.17}$$

When $\lambda = \mu$, (5.17) reduces to (5.13); when $\lambda = 1$ then $z = h$ and (5.17) reduces to (5.10); and when $\mu = 0$, (5.17) reduces to (5.12).

The results are summarized by

Theorem 5.20. *A horizontal slice of altitude $z \geq \mu h$ cut from a shell of altitude h and scaling factor μ has volume given by* (5.16)*, where $\lambda = z/h$. The altitude of its centroid above the base is given by* (5.17)*. In particular the formulas hold for any slice of a shell of an Archimedean, elliptic, or spherical dome.*

Theorem 5.16 covers the case $z \leq \mu h$.

In deriving the formulas in this section we made no essential use of the fact that the shells are elliptic. The important fact is that each shell is the region between two similar objects.

5.15 THE NECESSITY OF ELLIPTIC PROFILES

We know that every horizontal plane cuts an elliptic dome and its punctured cylindrical container in cross sections of equal area. This section reveals the surprising fact that the elliptical shape of the dome is actually a consequence of this property.

Consider a dome of altitude h, and its punctured prismatic counterpart having a congruent base bounded by a curve satisfying a polar equation $r = \rho(\theta)$. Each vertical half plane through the polar axis at angle θ cuts the dome along a curve we call a *profile*, illustrated by the example in Figure 5.20a. This is like the elliptic dome in Figure 5.17a, except that we do not assume that the profiles are elliptic.

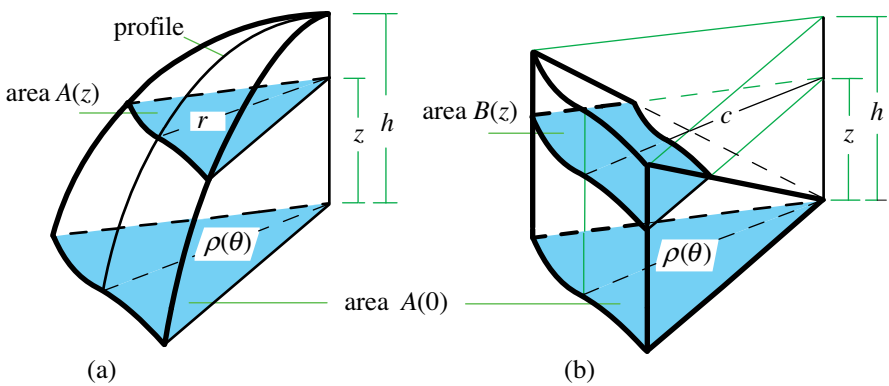


Figure 5.20: Determining the elliptic shape of the profiles as a consequence of the relation $A(z) = B(z)$.

Each profile passes through a point $(\rho(\theta), \theta)$ on the outer edge of the base. At an altitude z above the base a point on the profile is at a distance r from the polar axis, where r is a function of z that determines the shape of the profiles. We define a *general profile dome* to be one in which each horizontal cross section is similar to the base. Figure 5.20a shows a portion of a dome in which $\rho(\theta) > 0$. This portion

can be regarded as a wedge with two vertical plane faces that can be thought of as walls forming part of the boundary of the wedge.

Suppose that a horizontal plane at a distance z above the base cuts a region of area $A(z)$ from the wedge and a region of area $B(z)$ from the punctured prism. We know that $A(0) = B(0)$. Now we assume that $A(z) = B(z)$ for some $z > 0$ and deduce that the point on the profile with polar coordinates (r, θ, z) satisfies

$$\left(\frac{r}{\rho(\theta)}\right)^2 + \left(\frac{z}{h}\right)^2 = 1 \tag{5.18}$$

if $\rho(\theta) > 0$. In other words, the point on the profile at a height where the areas are equal lies on an ellipse with vertical semiaxis of length h, and horizontal semiaxis of length $\rho(\theta)$. Consequently, if $A(z) = B(z)$ for every z from 0 to h, the profile will fill out a quarter of an ellipse and the dome will necessarily be elliptic. Note that (5.18) implies that $r \to 0$ as $z \to h$.

To deduce (5.18), note that the horizontal cross section of area $A(z)$ in Figure 5.20a is similar to the base with similarity ratio $r/\rho(\theta)$, where $\rho(\theta)$ denotes the radial distance to the point where the profile intersects the base and r is the length of the radial segment at height z. By similarity, $A(z) = (r/\rho(\theta))^2 A(0)$. In Figure 5.20b, $B(z)$ is equal to $A(0)$ minus the area of a smaller similar region with similarity ratio $c/\rho(\theta)$, where c is the length of the parallel radial segment of the smaller similar region at height z. By similarity, $c/\rho(\theta) = z/h$, so

$$B(z) = (1 - (z/h)^2)A(0).$$

Equating this to $A(z)$ we find

$$(1 - (z/h)^2)A(0) = (r/\rho(\theta))^2 A(0),$$

which gives (5.18). And, of course, we already know that (5.18) implies $A(z) = B(z)$ for every z. Thus we have proved:

Theorem 5.21. *Corresponding horizontal cross sections of a general profile dome and its punctured prismatic counterpart have equal areas if, and only if, each profile is elliptic.*

As already remarked in Section 5.12, an elliptic dome can be deformed in such a way that areas of horizontal cross sections are preserved but the deformed dome no longer has elliptic profiles. At first glance, this may seem to contradict Theorem 5.21. However, such a deformation will distort the vertical walls; the dome will not satisfy the requirements of Theorem 5.21, and also the punctured counterpart will no longer be prismatic.

An immediate consequence of Theorem 5.21 is that any reducible general profile dome necessarily has elliptic profiles, because if all horizontal slices of such a dome and its counterpart have equal volumes then the cross sections must have equal areas. By a simple scaling argument, Theorem 5.21 can be extended to nonuniform general profile domes built from a finite number of general profile similar shells, each with its own constant density, under the condition that corresponding horizontal

slices of the dome and its counterpart have equal masses. For example, if the dome consists of two similar shells each with its own constant density, scale the outer shell with respect to the base of the polar axis so its new density is that of the inner shell. Do the same with its counterpart. This gives a new uniform dome and its counterpart whose horizontal cross sections have equal areas. Hence by Theorem 5.21 every profile of the new dome is elliptic, so the original dome also has elliptic profiles. By induction we obtain the result for a dome built from any number of similar shells.

NOTES ON CHAPTER 5

Most of this chapter is a compilation of work published in [15] and [20], the first of which was awarded a Lester R. Ford Award in 2005. The motivation for this chapter was to extend to more general solids classical properties that seemed to be unique to spheres and hemispheres. Initially an extension was given for Archimedean domes, and a further extension was made by simply dilating the domes in a vertical direction. The extensions could also have been analyzed by using properties of inscribed spheroids. A significant extension was made when we introduced polygonal elliptic domes whose bases could be arbitrary polygons, not necessarily circumscribing the circle. In this case there are no inscribed spheroids to aid in the analysis, but the method of punctured containers was applicable. This led naturally to general elliptic domes with arbitrary base, and the method of punctured containers was formulated in terms of mappings that preserve volumes. But the real power of the method is revealed by the treatment of nonuniform mass distributions. Problems of determining volumes and centroids of elliptic wedges, shells, and their slices, including those with cavities, were reduced to those of simpler prismatic containers. Finally, we showed that domes with elliptic profiles are essentially the only ones that are reducible.

Animated versions of parts of this chapter can be viewed at

http://www.its.caltech.edu/~mamikon/globes.html .

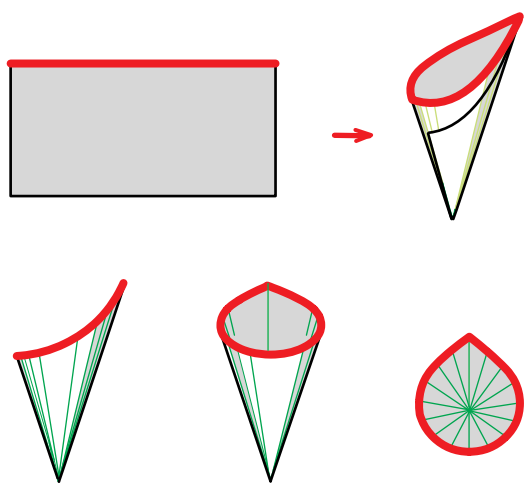
For an intriguing story regarding the oldest known surviving copy of the work of Archimedes see [43], or search the Internet for *Archimedes Palimpsest.*

Chapter 6

UNWRAPPING CURVES FROM CYLINDERS AND CONES

This problem can be easily solved by the methods developed in this chapter. The reader may wish to try solving it before reading the chapter.

A rectangular sheet of paper is wrapped to form part of the surface of a circular cone as shown. The top edge of the rectangle forms the curved edge of the surface. Various views of the curved edge of the surface are shown below.



(a) side view (b) front view (c) top view

Give an analytic description, such as a cartesian equation, a polar equation, or parametric equations for each of the curves in (a), (b), *and* (c). *Consider also the special case in which the vertex angle of the cone is* $\pi/3$.

CONTENTS

PART 1: UNWRAPPING FROM CYLINDERS

PART 2: UNWRAPPING FROM CONES

6

What happens to the shape of a curve lying on the surface of a circular cylinder when the cylinder is unwrapped onto a plane? Conversely, draw a plane curve on transparent plastic, and roll it into cylinders of different radii. What shapes does the curve take on the cylinders? How do they appear when viewed from different directions? Similar questions are investigated for space curves unwrapped from the surface of a right circular cone, including conic sections, spirals, and geodesics. Unwrapped conic sections produce a new class of plane curves called generalized conics. This chapter formulates these somewhat vague questions in terms of equations, and analyzes them with surprisingly simple two-dimensional geometric transformations that lead to many unexpected results. The methods for analyzing cones and cylinders differ substantially, but both use the fact that unwrapping a developable surface preserves area and arclength. Applications are given to diverse fields such as descriptive geometry, computer graphics, sheet metal construction, and educational hands-on activities.

PART 1: UNWRAPPING FROM CYLINDERS

6.1 INTRODUCTION

In his delightful book *Mathematical Snapshots*, Steinhaus [66] describes a simple, engaging construction, illustrated in Figure 6.1. Wrap a piece of paper around a cylindrical candle, and cut it obliquely with a knife. The cross section is an ellipse, which becomes a sinusoidal curve when unwrapped.

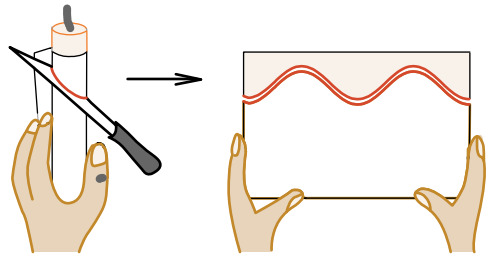


Figure 6.1: Elliptical cross section of a cylinder becomes sinusoidal when unwrapped.

The same idea can be demonstrated with a safer instrument. Take a cylindrical paint roller, dip it at an angle in a container of paint or water color, and roll it on a flat surface. The roller prints a sinusoidal wave pattern as shown in Figure 6.2.

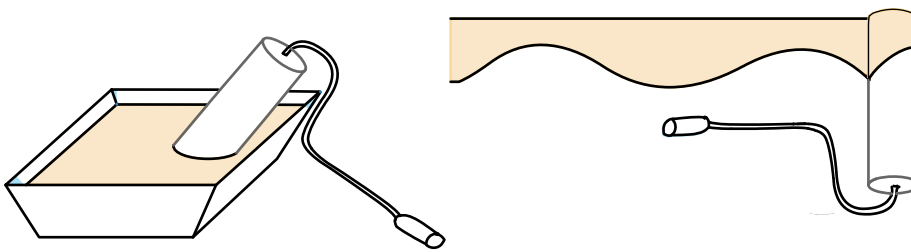


Figure 6.2: A paint roller used to print sinusoidal waves on a flat surface.

Now imagine the elliptical cross section replaced by any curve lying on the surface of a right circular cylinder. What happens to this curve when the cylinder is unwrapped?

Consider also the inverse problem, which you can experiment with by yourself: Start with a plane curve (line, circle, parabola, sine curve, etc.) drawn with a felt pen on a rectangular sheet of transparent plastic, and roll the sheet into cylinders of different radii. What shapes does the curve take on the cylinders? How do they appear when viewed from different directions? A few trials reveal an enormous number of possibilities, even for the simple case of a circle.

Part 1 of this chapter formulates these somewhat vague questions more precisely, in terms of equations, and shows that they can be answered with surprisingly simple 2-dimensional geometric transformations, even when the cylinder is not circular. For a circular cylinder, a sinusoidal influence is always present, as exhibited above. And we demonstrate, through examples, applications to diverse fields such as descriptive geometry, computer graphics, printing, sheet metal construction, and educational hands-on activities. Part 2, beginning with Section 6.12, treats curves unwrapped from the surface of a right circular cone.

6.2 UNWRAPPING AN ELLIPSE FROM A CIRCULAR CYLINDER

Before turning to the general problem, let's analyze the foregoing sinusoidal construction. Cut a right circular cylinder of radius r by a plane through a diameter of its base, at angle of inclination β, where $0 < \beta < \pi/2$. The example in Figure 6.3a shows one-half of the elliptical cross section and a wedge cut from the cylinder. A vertical cutting plane parallel to the major axis of the ellipse intersects the wedge along a right triangle T (shown shaded) with base angle β.

When the lateral surface of the cylinder is unwrapped onto a plane, the circular base unfolds along a line we call the x axis. Here x is the length of the circular arc measured from A at the extremity of the base diameter, to B at the base of triangle T, as shown in Figure 6.3a. The base of T has length $r\sin(x/r)$, and its height is equal to $h\sin(x/r)$, where $h = r\tan\beta$, so the unwrapped curve is the graph of

$$u(x) = h\sin\frac{x}{r}$$

representing a sinusoidal curve with period $2\pi r$ and amplitude h. For fixed r, the amplitude h increases with β.

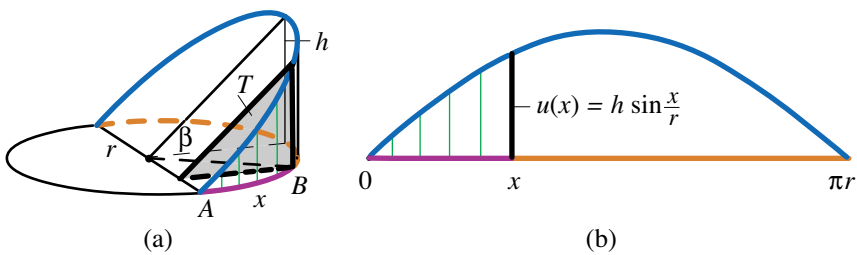


Figure 6.3: The circular arc AB of length x in (a) unwraps onto the line segment $[0, x]$ in (b). The altitude of triangle T unwraps onto the height $u(x)$.

An adjacent arch below the x axis comes from unwrapping the second symmetrically located wedge. Volume calculations of cylindrical wedges like these were considered by Archimedes and are analyzed in more detail in Chapter 5, where unwrapping of a cylinder is also used to deduce the quadrature of a sine curve without integral calculus (see Section 5.8).

6.3 CURVE OF INTERSECTION OF TWO CYLINDERS

A *cylinder* is any surface generated or swept out by a straight line moving along a plane curve and remaining parallel to a given line. The curve is called a *directrix* of the cylinder and the moving line that sweeps out the cylinder is called a *generator*. The directrix is not unique because any plane cuts a given cylinder along a plane curve that can serve as directrix. When the cutting plane is perpendicular to the generators we call the directrix a *profile* of the cylinder.

A curve in the xy plane has an implicit cartesian equation of the form

$$m(x, y) = 0.$$

In xyz space the equation describes a cylinder having this profile, with generators parallel to the z axis. Similarly, an equation of the form $p(x, z) = 0$ (with y missing) describes a cylinder with generators parallel to the y axis, whereas one of the form $q(y, z) = 0$ (with x missing) describes a cylinder with generators parallel to the x axis.

Start with a vertical cylinder in xyz space with equation $m(x, y) = 0$, which we call the *main cylinder*, and locate the coordinate axes so that the z axis lies along a generator and the x axis is tangent to the profile. Intersect the main cylinder with a horizontal cylinder $p(x, z) = 0$, which we call the *cutting cylinder*. Their curve of intersection C is the set of points (x, y, z) that satisfy both $m(x, y) = 0$ and $p(x, z) = 0$. Let C_p denote the profile of the cutting cylinder, which shows what C looks like when viewed along the generators of the cutting cylinder (see Figure 6.4a). We call the xz plane the *viewing plane*, and the equation $p(x, z) = 0$ the *profile equation*.

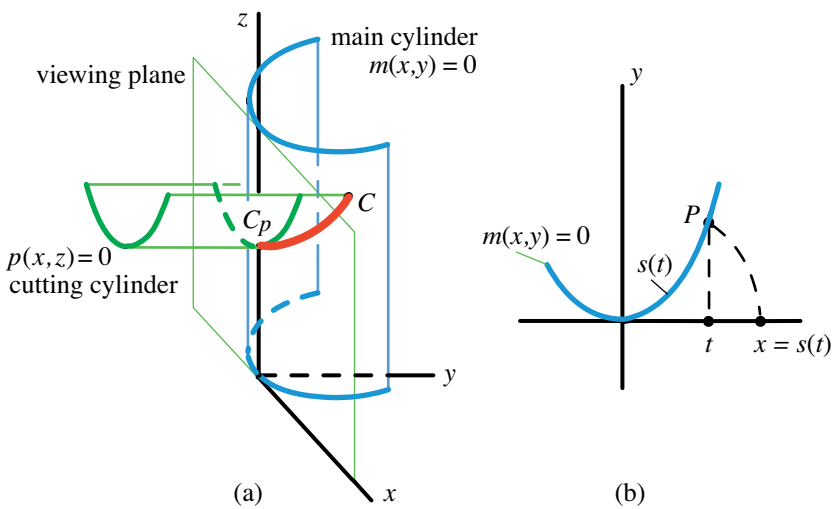


Figure 6.4: Curve of intersection of a main cylinder and an orthogonal cutting cylinder.

6.4 UNWRAPPING A CURVE FROM ANY CYLINDER

Now unwrap the main cylinder onto the xz plane, which we call the *unwrapping plane*. A curve C on the cylinder is printed onto an unwrapped curve C_u in the xz plane. It has an equation of the form $u(x, z) = 0$, called an *unwrapping equation*, which we shall determine from the profile equations $m(x, y) = 0$ and $p(t, z) = 0$ that define C.

We use the fact that every cylinder is a developable surface, hence unwrapping preserves distances. In particular, any arc of length s on the (horizontal) profile $m(x, y) = 0$ gets printed onto a line segment of the same length on the x axis.

To formulate this in terms of equations, imagine each point P on the profile of the main cylinder described in terms of new parameters t and $s(t)$, where $s(t)$ is the arclength of the portion of the profile joining the origin to P, and t is the projection of that arc on the x axis. Unwrapping the cylinder prints P onto a point on the xz plane with coordinates $(s(t), 0)$ (Figure 6.4b). Hence any other point on C at a height z above P is printed onto the point $(s(t), z)$, where z satisfies the profile equation $p(t, z) = 0$. Consequently, in the unwrapping equation $u(x, z) = 0$, x and z are related by $x = s(t)$ and $p(t, z) = 0$. We plot t on the x axis.

To express u directly in terms of p, we consider portions of C for which the function $x = s(t)$ has an inverse, so that t can be expressed in terms of x by the relation $t = s^{-1}(x)$. Under these conditions, we have the following theorem.

Theorem 6.1. *For an intersection curve C as described above, the profile equation $p(t, z) = 0$ of C_p and the unwrapping equation $u(x, z) = 0$ of C_u are related by*

$$u(x, z) = p(s^{-1}(x), z), \tag{6.1}$$

$$p(t, z) = u(s(t), z). \tag{6.2}$$

When curves C_p and C_u are described by explicit equations, we use the same letters p (for profile) and u (for unwrapping) and write $z = p(t)$ and $z = u(x)$, respectively. In this case (6.1) and (6.2) become

$$u(x) = p(s^{-1}(x)), \quad p(t) = u(s(t)).$$

In other words, to obtain the profile function $p(t)$ from $u(x)$, replace the argument x by the arclength $s(t)$. And, conversely, to obtain the unwrapping function $u(x)$ from $p(t)$, replace the argument t by the inverse $s^{-1}(x)$.

The following special case is worth noting.

Corollary 6.1. *If the unwrapping function is linear, $u(x) = x$, then the profile function is the arclength function: $p(t) = s(t)$. And if the profile function is linear, $p(t) = t$, then the unwrapping function is the inverse of the same arclength function: $u(x) = s^{-1}(x)$.*

6.5 UNWRAPPING A CURVE FROM A CIRCULAR CYLINDER

Figure 6.5 illustrates these ideas when the main cylinder is a right circular cylinder of radius r. In Figure 6.5c, a circular arc of length $r\theta$ is unwrapped onto a segment of length x, and its horizontal projection has length $t = r \sin\theta$. Because $\theta = x/r$, the arclength function is $x = s(t) = r \arcsin(t/r)$, and its inverse is $t = s^{-1}(x) = r \sin(x/r)$.

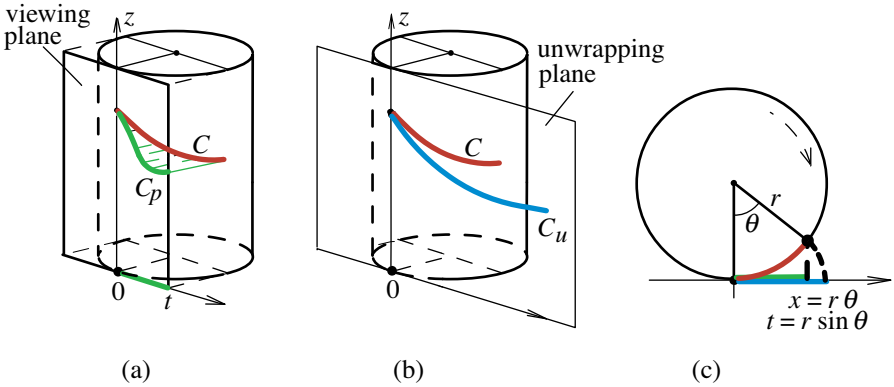


Figure 6.5: (a) Curve C on a circular cylinder and its horizontal profile C_p on a tangent viewing plane. (b) The unwrapped curve C_u obtained by rolling the cylinder along the unwrapping plane. (c) A point on a circle projects onto $t = r\sin\theta$, but unwraps onto $x = r\theta$.

Thus, Theorem 6.1 becomes:

Theorem 6.2. *On a right circular cylinder of radius r, let C be a curve defined by the profile C_p of a horizontal cutting cylinder, and let C_u denote its unwrapped image. Then the profile equation $p(t, z) = 0$ for C_p and the unwrapping equation $u(x, z) = 0$ for C_u are related by*

$$u(x, z) = p(r \sin \frac{x}{r}, z), \tag{6.3}$$

$$p(t, z) = u(r \arcsin \frac{t}{r}, z). \tag{6.4}$$

This shows that the sine function is always present when a curve is unwrapped from a right circular cylinder onto a plane. When C_p and C_u are described by explicit equations, say $z = p(t)$ and $z = u(x)$, then (6.3) and (6.4) become

$$u(x) = p(r \sin \frac{x}{r}), \tag{6.5}$$

$$p(t) = u(r \arcsin \frac{t}{r}). \tag{6.6}$$

If $r = 1$, then $\arcsin t = s(t)$, the length of the circular arc whose sine is t.

Now we apply Theorem 6.2 to some simple examples. The first two illustrate Corollary 6.1.

Example 1 (Linear profile function $p(t)=ct$). In this case, the cutting cylinder is a plane through the line $z = ct$, where c is constant. From (6.5) we find that the unwrapping function is $u(x) = cr\sin(x/r)$, whose graph is a sinusoidal curve with period $2\pi r$.

If the cutting plane is inclined at an angle β with a horizontal diameter of the cylinder, then $c = \tan\beta$ and the unwrapping function is $u(x) = h\sin(x/r)$, where $h = r\tan\beta$, in agreement with the result obtained earlier by analyzing Figure 6.3. The shape of the cross section curve C itself depends on the direction from which it is viewed. When viewed along the edge of the cutting plane we see the profile C_p as a line segment. In a later example we show that when viewed from any direction the cross section cut by a plane is, as expected, always an ellipse (possibly degenerate).

Example 2 (Linear unwrapping function $u(x)=cx$). This example explains what happens when a straight line on a transparency is rolled onto a cylinder of radius r. The corresponding profile function obtained from (6.6) is

$$p(t) \;=\; cr\arcsin\frac{t}{r}.$$

Because distances are preserved when the cylinder is unwrapped, a line segment on the unwrapped cylinder (the shortest path between its endpoints) becomes a geodesic arc (the shortest path) on the cylinder, no matter how tightly it is rolled. In other words, on a right circular cylinder the profile of a geodesic arc is part of an arcsine curve.

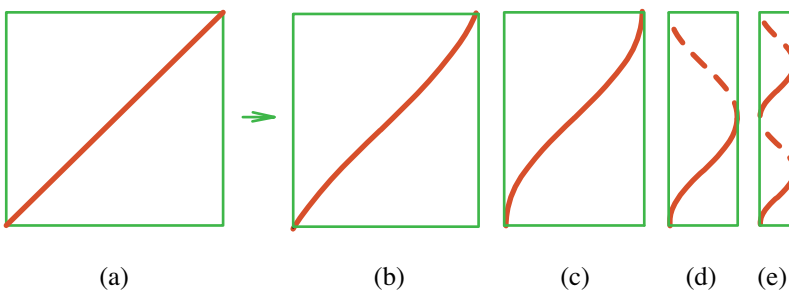


Figure 6.6: A line segment (a) wraps onto a geodesic. (b)-(e) Geodesic profiles (arcsine curves) on cylinders of decreasing radii. The dashed curves in (d) and (e) are on the rear half of the cylinder.

Figure 6.6a shows the line $u(x) = x$ in the unwrapping plane, and Figures 6.6b-e show the profile of the geodesic arc on cylinders of decreasing radii. Again we see sinusoidal curves, but they are flipped sideways, as predicted by Corollary 6.1. The dashed curve in Figure 6.6d indicates the portion of the geodesic arc that lies on the rear half of the cylinder. The two repeated curves in Figure 6.6e represent different branches of the arcsine function.

This suggests a simple educational hands-on activity that can engage young students while they learn that a geodesic on a circular cylinder is always part of a circular helix. Use a felt pen to draw a line segment on a transparency, roll it into a circular cylinder, view the profile in various directions, and watch the sine waves change shape as the radius of the cylinder varies.

Example 3 (Parabolic cutting cylinder). Figure 6.7 shows a quadratic profile equation $p(t) = ct^2$ for a constant $c > 0$. By (6.5) the corresponding unwrapping function for C_u is

$$u(x) = cr^2 \sin^2 \frac{x}{r}. \tag{6.7}$$

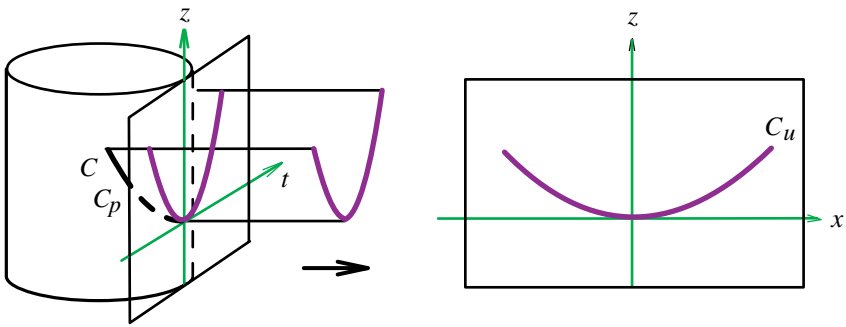


Figure 6.7: Curve C cut by parabolic cylinder, with profile C_p and unwrapped curve C_u.

Examples of this type have practical applications. To illustrate, take a rectangular piece of sheet metal cut along the curve described by (6.7), and roll it to form a circular cylinder of radius r. The curve C cut out on the resulting cylinder indicates exactly where it will intersect a parabolic gutter having profile $p(t) = ct^2$.

Example 4 (Wrapping a circle onto a cylinder). On a sheet of transparent plastic, draw a unit circle, and roll the sheet into a circular main cylinder. What does the wrapped circle look like when viewed from the side? The circle wraps onto a curve C and we want its horizontal profile C_p. The upper half of the unwrapped unit circle can be described by the unwrapping function

$$u_+(x) = \sqrt{1 - x^2},$$

and the lower half by $u_-(x) = -\sqrt{1 - x^2}$, shown dotted in Figure 6.8a. Both halves can be described by the implicit equation

$$u^2(x) \;=\; 1 - x^2.$$

From (6.5) we see that the corresponding profile functions are $p_\pm(t) = u_\pm(r \arcsin \frac{t}{r})$, both of which can be described by the implicit equation

$$p^2(t) = 1 - (r \arcsin \frac{t}{r})^2.$$

They depend on the radius of the main cylinder.

In Figures 6.8b-f the cylinder is turned (for ease in displaying) so that its axis is horizontal, and the corresponding graph of $p_+(t)$ is shown for various values of

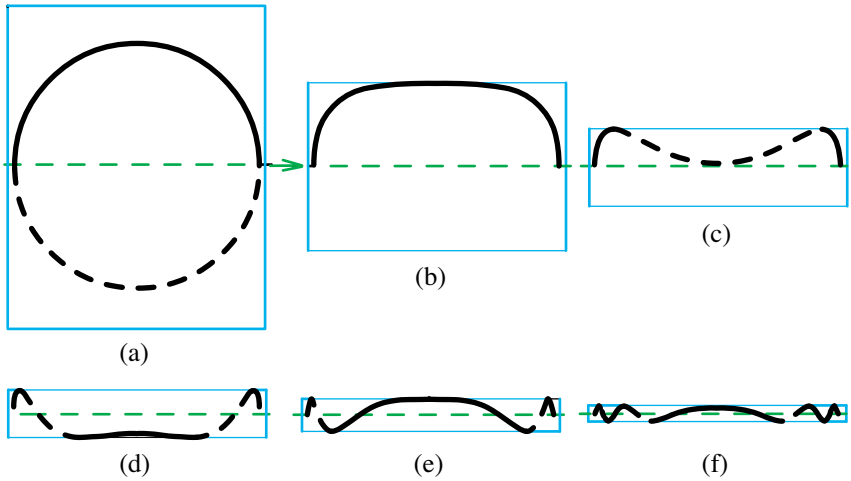


Figure 6.8: Circle on a transparency rolled onto cylinders of decreasing radii.

radius r. The flipped graph of each $p_-(t)$ (not shown) is the mirror image reflection through the horizontal dashed line. In Figures 6.8c-f the dashed curves lie on the rear half of the cylinder and are hidden from view.

6.6 ROTATING THE MAIN CYLINDER

On a main circular cylinder of radius r, take a curve C with explicit profile function $z = p(t)$. Rotate the cylinder through an angle α about its axis, but keep the viewing plane fixed. The profile function of the rotated curve on the viewing plane depends on α and we denote its ordinates by z_α. The next theorem describes z_α in terms of p.

Theorem 6.3. *On a cylinder of radius r, take a curve C with profile function $z = p(t)$ on the viewing plane. If the cylinder is rotated about its axis through an angle α, the rotated curve on the same viewing plane has profile function*

$$z_\alpha = p(t\,\cos\alpha + \sqrt{r^2 - t^2}\,\sin\alpha). \tag{6.8}$$

Proof. Rotation of the cylinder through an angle α is equivalent to shifting the arclength $x = r\theta$ by an amount $r\alpha$. Therefore, if the unwrapping function of C is $u(x)$, rotation of the cylinder through an angle α (measured clockwise when viewed from above) replaces x by $x + r\alpha$, and the unwrapping equation of rotated C becomes

$$z_\alpha = u(x + r\alpha) = p(r\,\sin\frac{x + r\alpha}{r})$$

by (6.5). But

$$r\,\sin(\frac{x}{r} + \alpha) = r\,\sin\frac{x}{r}\,\cos\alpha + r\cos\frac{x}{r}\,\sin\alpha.$$

In terms of $t = r\sin(x/r)$, we have $r\cos(x/r) = \sqrt{r^2 - t^2}$, and the foregoing equation for z_α becomes

$$z_\alpha = p(\, r\, \sin\frac{x + r\alpha}{r}) = p(t\, \cos\alpha + \sqrt{r^2 - t^2}\sin\alpha),$$

which proves (6.8). Note that $z_0 = p(t)$.

It is not surprising that the combination $t\cos\alpha + \sqrt{r^2 - t^2}\sin\alpha$ in (6.8) resembles the right side of the equation $x' = x\cos\alpha + y\sin\alpha$ for changing coordinates from an xy system to an $x'y'$ system by rotation of axes through an angle α.

We leave it to the reader to formulate the result corresponding to (6.8) when the profile is given in implicit form $p(z, t) = 0$.

Corollary 6.2. (Perpendicular view) *When $\alpha = \pi/2$, we get the profile function*

$$z_{\pi/2} = p(\sqrt{r^2 - t^2}).$$

Example 5 (Rotated view of a slanted plane cut). If $p(t) = t$, which corresponds to cutting the original cylinder by a plane inclined at 45°, then (6.8) becomes

$$z_\alpha = t\, \cos\alpha + \sqrt{r^2 - t^2}\sin\alpha,$$

which implies

$$z_\alpha^2 - 2z_\alpha t\, \cos\alpha + t^2 = r^2\, \sin^2\alpha.$$

As expected, this represents an ellipse (possibly degenerate) in the tz_α plane. Examples are shown in Figure 6.9. When $\alpha = 0$ the profile is a line segment, $z = t$, and when $\alpha = \pi/2$, it is a circle, $z^2 + t^2 = r^2$.

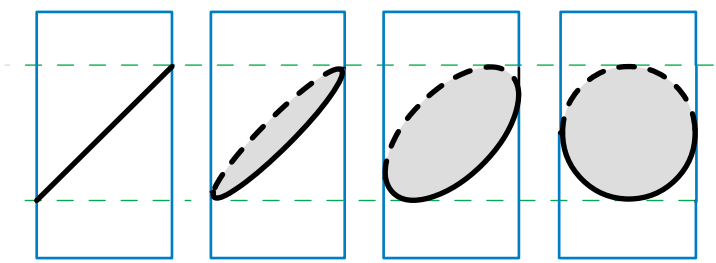


Figure 6.9: Various profiles of a rotating inclined ellipse. The dashed portions lie on the rear half.

Example 6 (Rotated view of a parabolic cut). In this case $p(t) = ct^2$ and (6.8) becomes

$$z_\alpha = c(t\, \cos\alpha + \sqrt{r^2 - t^2}\sin\alpha)^2.$$

For $c = 1$, Figure 6.10 shows the profile curve C_p for various values of α. Surprisingly, when the cylinder is rotated through a right angle, the profile is the mirror image of the original parabola reflected through the line $z = r^2/2$.

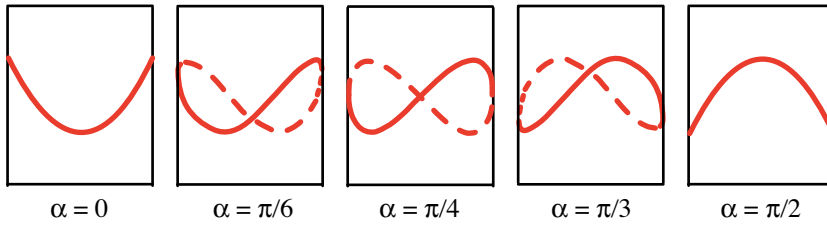


Figure 6.10: Rotated views of a parabolic intersection. In a perpendicular direction we see the original parabola flipped upside down.

6.7 CYLINDER TO CYLINDER

Relations (6.5) and (6.6) can also be used to analyze what happens when a curve on one circular cylinder is unwrapped onto another cylinder and viewed on a common viewing plane, all three tangent along a common generator. Start with a curve $C(r)$ on a right circular cylinder of radius r with profile function p_r, and first unwrap it onto a plane curve C_u with unwrapping function $z = u(x)$, where

$$u(x) = p_r(r \sin \frac{x}{r}).$$

Now take the unwrapped curve C_u and let $C(R)$ be the curve obtained by wrapping C_u back onto a right circular cylinder of radius R. Because $C(R)$ has the same unwrapping function u, its profile function p_R satisfies

$$u(x) = p_R(R \sin \frac{x}{R}).$$

Now equate the two expressions for $u(x)$ to obtain:

Theorem 6.4. *The profile functions p_r and p_R of a curve on two tangent cylinders of respective radii r and R are related by*

$$p_r(r \sin \frac{x}{r}) = p_R(R \sin \frac{x}{R}). \tag{6.9}$$

In particular, if one cylinder has radius $r = 1$, we can use (6.9) to see how the profile of $C(R)$ varies with R. In this case, p_R and p_1 satisfy

$$p_R(R \sin \frac{x}{R}) = p_1(\sin x). \tag{6.10}$$

Equation (6.10) can also be written as

$$p_R(t) = p_1(\sin(R \arcsin \frac{t}{R})).$$

The dependence on R can be regarded as a movie that shows various stages of the process of unwrapping. As $R \to \infty$ the cylinder of radius R becomes the tangent viewing plane and (6.10) becomes $u(x) = p_1(\sin x)$, in agreement with (6.5).

Example 7 (Unwrapping an ellipse from one cylinder to another). Take $p_1(t) = t$,which corresponds to cutting the main cylinder of radius 1 by a plane inclined at 45°. The curve of intersection on the original cylinder is an ellipse C but its profile on the viewing plane appears as a line segment. When we unwrap the unit cylinder onto a cylinder of radius R, the profile function p_R of the ellipse becomes, according to (6.11),

$$p_R(t) = \sin(R \arcsin \frac{t}{R}).$$

Figure 6.11 shows the profile of curve $C(R)$ for various values of R, starting with $R = 1$ in Figure 6.11c. Smaller radii are shown in Figures 6.11a and b, and larger radii in Figures 6.11d and 6.11e. The limiting case $R \to \infty$ is a pure sine curve, $z = \sin t$.

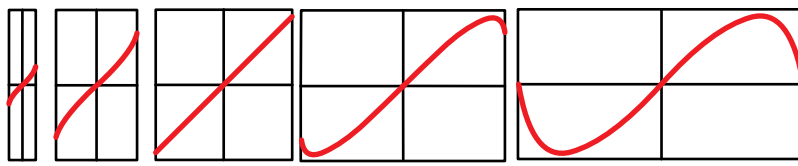


Figure 6.11: The shape of the profile curve varies with the radius of the cylinder, from a line segment in (c) ($R = 1$), to smaller radii in (a), (b) and larger radii in (d), (e).

6.8 DRILLED CYLINDER

Through the main cylinder of radius r, drill a hole with a circular cylindrical drill of radius a whose axis is perpendicular to the viewing plane at a distance d from the axis of the cylinder, where $0 \le d \le r + a$. The edge of the hole is a curve C on the given cylinder that appears as part of a circle when viewed along the axis of the drill. A plane through the axis of the main cylinder parallel to the viewing plane divides the cylinder into two parts, a front portion and a mirror image rear portion. The curve C will also consist of two parts, one lying on the front portion, and the mirror image on the rear portion. The two parts can be connected or disconnected, depending on the size and position of the hole. Also, the corresponding unwrapped curve C_u will be symmetric about the vertical line $x = \pi r/2$.

Place the axis of the drill so it intersects the t axis of the viewing plane orthogonally at $(d, 0)$. To find the unwrapping function $z = u(x)$ of C, first we find the profile function $z = p(t)$, then by (6.5) we have $u(x) = p(r \sin(x/r))$. When $r = 1$, we have $u(x) = p(\sin x)$.

To determine $p(t)$, note that each projected point $(t, p(t))$ in the viewing plane lies on a circle of radius a with center at $(d, 0)$, so $p(t)^2 + (d - t)^2 = a^2$. Hence the

upper and lower halves of the circular hole have profile functions that satisfy

$$p^2(t) = a^2 - (d-t)^2\,. \tag{6.11}$$

The corresponding unwrapping functions satisfy

$$u^2(x) = a^2 - (d - r\,\sin\frac{x}{r})^2. \tag{6.12}$$

The following examples display interesting families of unwrapped curves obtained when a, d, and r are treated as parameters.

Example 8 (Drill of same radius as main cylinder; variable distance *d*). Take $a = r = 1$ in (6.12), and let d decrease from 2 to 0. When $d = 2$, the drill is tangent to the main cylinder at one point, which unwraps onto the single point $(\pi/2, 0)$. For $d = 1$, the unwrapping equation is

$$u^2(x) = \,2\sin x - \sin^2 x\,.$$

As d decreases, the hole changes shape, reaching its maximum size when $d = 0$, at which stage the drill's axis passes through the axis of the main cylinder and

$$u^2(x) = 1 - \sin^2 x \;= \cos^2 x.$$

Figure 6.12 shows the upper half of the unwrapped curve for decreasing values of d.

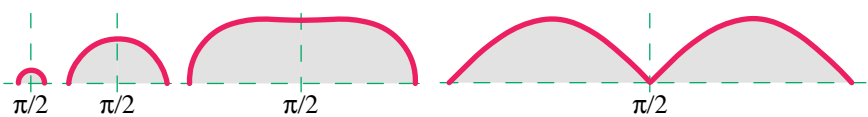


Figure 6.12: The unwrapped curve (upper half) obtained by drilling a hole of unit radius through different parts of the main cylinder of the same radius. Each is symmetric about the line $x = \pi/2$. The lower half (not shown) is the reflection of the upper half through the x axis.

Each curve shown in Figure 6.12 includes the unwrapped symmetric image that comes from the rear portion of the main cylinder, with vertical axis of symmetry $x = \pi/2$. In each case the lower half (not shown) can be obtained by reflecting the curve through the x axis. Incidentally, the graph of $z = |\cos x|$, together with its reflection, $z = -|\cos x|$, for $|x| \le \pi/2$, represent the unwrapped intersection of two perpendicular cylinders of unit radius. In this case the intersection itself is an ellipse (see [2]).

Example 9 (Centered drill of variable radius). If $r = 1$ and $d = 0$, the hole is centered on the axis of the main cylinder, and (6.12) becomes $u^2(x) = a^2 - \sin^2 x$, which represents a family of unwrapped curves depending on the radius a of the hole. In this example the geometry and the equation itself show that each unwrapped curve is symmetric about the line $x = 0$. Figure 6.13 shows a few members of the family, $z = \sqrt{a^2 - \sin^2 x}$, for increasing values of a, from very small radius to very large radius. The case $a = 1$ gives the third curve, $z = |\cos x|$, also in Figure 6.12.

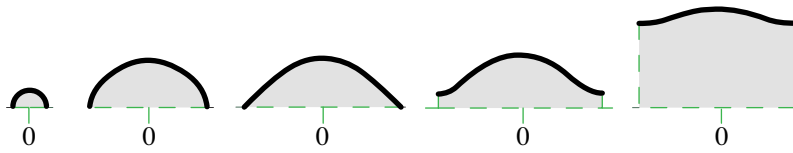


Figure 6.13: The unwrapped image of the upper half of a hole of variable radius drilled through the axis of the main cylinder. Each is symmetric about the line $x = 0$. The lower half (not shown) is the reflection of the upper half through the x axis.

Example 10 (Rotated view of a drilled circular hole). Return to (6.11), which describes the projected view of the hole obtained by drilling through the main cylinder of radius r with a drill of radius a whose axis is perpendicular to the viewing plane at a distance d from the axis of the cylinder, where $0 \le d \le r + a$. Let's find the shape of the hole as projected in the viewing plane after the main cylinder has been rotated through a right angle. Applying the general rotation formula in (6.8) with $\alpha = \pi/2$ and $p(t)$ as given in (6.11), we find the upper and lower halves of the hole in the rotated cylinder satisfy

$$z^2 = a^2 - (d - \sqrt{r^2 - t^2})^2 \,. \tag{6.13}$$

This can be written as

$$z^2 = a^2 - d^2 - r^2 + t^2 + 2d\sqrt{r^2 - t^2},$$

or

$$(z^2 - a^2 + d^2 + r^2 - t^2)^2 = 4d^2(r^2 - t^2),$$

a cartesian equation of degree 4 in t and in z.

When $d = r+a$, the drill is tangent to the main cylinder. As d decreases towards 0 the projection of the rotated hole changes its appearance. When $d = 0$, the drill passes through the axis of the cylinder and the cartesian equation simplifies to

$$t^2 - z^2 = r^2 - a^2,$$

which represents an equilateral hyperbola if $a \neq r$. The hyperbola has a horizontal axis if $a < r$ and a vertical axis if $a > r$. If $a = r$, the radius of the drill is the same as that of the cylinder and the rotated projected curves are the pair of lines $z = \pm t$.

Figure 6.14 shows how the projection changes its appearance when $r = a = 1$, and the axis of the drill moves toward the axis of the main cylinder. In Figures 6.14a-d it has the appearance of an expanding oval. In Figure 6.14e it becomes nonconvex, then gradually deforms to resemble hyperbolas (Figure 6.14f). Finally the two axes intersect when $d = 0$ and it becomes a pair of lines, a degenerate hyperbola (Figure 6.14g).

Figure 6.15 shows a corresponding sequence when $r = 1$ and $a = 1/2$. When $d = 0$ an equilateral hyperbola suddenly appears. When $a > 1$ the projections are like those in Figure 6.15, but turned sideways by 90^o.

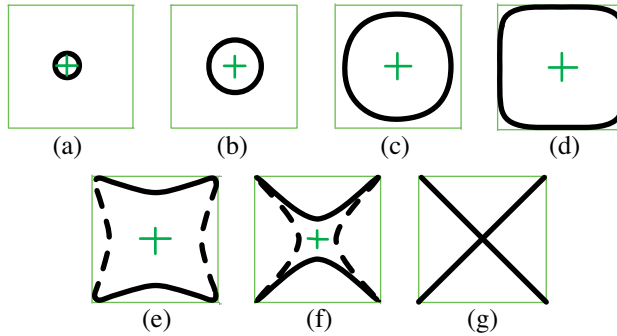


Figure 6.14: Profiles of the hole made by a drill of radius equal to that of the main cylinder, as the axis of the drill moves toward the axis of the main cylinder, viewed from a direction perpendicular to the axis of the drill.

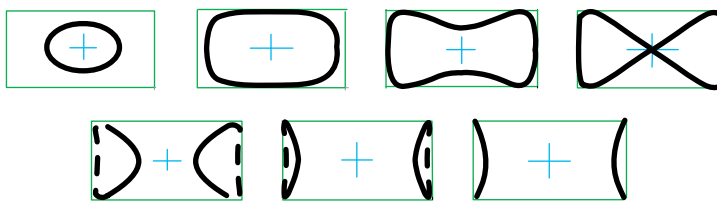


Figure 6.15: Profiles of the hole made by a drill of radius half that of the main cylinder, as the axis of the drill moves toward the axis of the main cylinder, viewed from a direction perpendicular to the axis of the drill.

When $\sqrt{r^2 - t^2}$ is replaced with $\sqrt{r^2 + t^2}$ in (6.13), the resulting equation describes a cross section of a torus.

Figure 6.16 shows some curves of intersection of a torus with a vertical plane, and provides an interesting analogy to Figure 6.15.

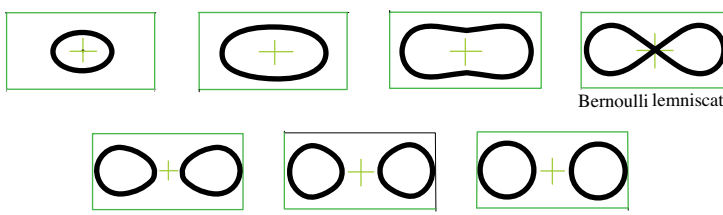


Figure 6.16: Curves of intersection of a torus with a vertical plane.

6.9 TILTED CUTTING CYLINDER

Up to now we have studied the profile of a curve C cut from the main cylinder by an orthogonal cutting cylinder. In descriptive geometry and in applications to sheet metal work the cutting cylinder is not always orthogonal to the main cylinder but may be tilted at an angle β. For the applications, we want to know what the unwrapped version C_u looks like so we can cut the unwrapped cylinder along this curve. To find C_u it suffices to determine the horizontal projection C_p, which is related to the profile of the slanted cutting cylinder. Figure 6.17 reveals this relation.

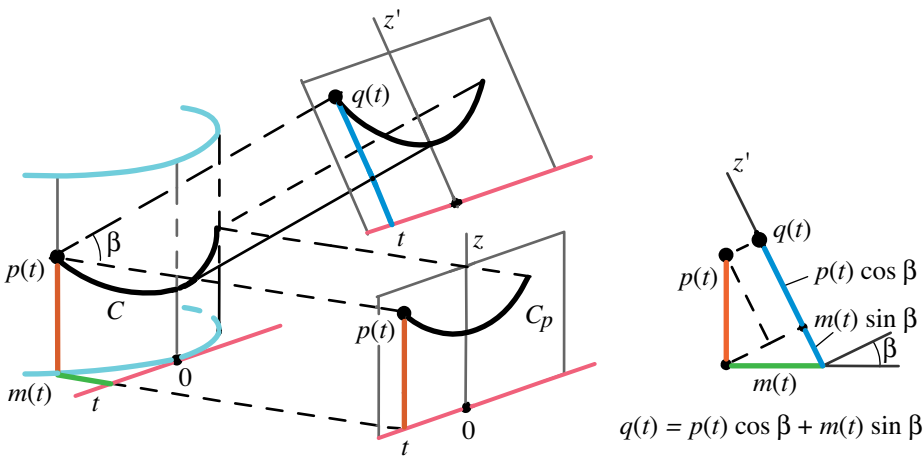


Figure 6.17: Main cylinder cut by a tilted cylinder, and a horizontal view along the t axis.

Assume the main cylinder has directrix with explicit equation $y = m(t)$ in the horizontal ty plane. Introduce a z' axis perpendicular to the horizontal t axis and making an angle β with the vertical z axis. Let $z' = q(t)$ denote the equation in the tz' plane of the profile of the tilted cutting cylinder. (Although $m(0) = 0$, we do not require that $q(0) = 0$.) The relation between $q(t)$ and the horizontal projection function $z = p(t)$ is given in (6.14). If $q(t)$ is known, this determines $p(t)$ which leads to the required unwrapping function $u(x)$.

Theorem 6.5. *Let C be the curve of intersection of a main cylinder with horizontal profile $y = m(t)$, cut by another cylinder tilted at an angle β, with profile function $z' = q(t)$. Then the horizontal projection $z = p(t)$ of C_p is related to the function $z' = q(t)$ by*

$$q(t) = p(t)\,\cos\,\beta\, + m(t)\,\sin\,\beta. \tag{6.14}$$

Figure 6.17 provides a geometric proof of (6.14), which represents a rotation from the z axis to the z' axis. The term $p(t)\cos\beta$ is the projection of $p(t)$ onto the tilted viewing plane, and the term $m(t)\sin\beta$ is an upward shift due to the angle of

view. In particular, if the main cylinder is circular of radius r with directrix tangent to the t axis, then $(y-r)^2+t^2=r^2$, so $y=m(t)=r-\sqrt{r^2-t^2}$.

Special cases.

(a) If $\beta=0$, this gives $q(t)=p(t)$.

(b) If $\beta=\pi/2$, (6.14) becomes $z'=m(t)$. This is to be expected because the viewing direction is along the axis of the main cylinder, and every curve on the main cylinder appears as part of its profile.

(c) If $p(t)=0$, (6.14) becomes $z'=m(t)\sin\beta$, a scaled version of the profile of the main cylinder. In particular, if the main cylinder has a circular profile $m(t)=r-\sqrt{r^2-t^2}$, then for $\beta\neq 0$ the relation $z'=m(t)\sin\beta$ can also be written as

$$\frac{t^2}{r^2}+\left(\frac{z'-r\sin\beta}{r\sin\beta}\right)^2=1.$$

This is the equation of an ellipse with center at $(0,\ r\sin\beta)$ in the tz' plane and with semiaxes of lengths r and $r\sin\beta$. In this case, C is a circle of radius r on the main cylinder, and it appears as an ellipse when projected on a tilted plane.

(d) If the main cylinder has a circular profile $m(t)=r-\sqrt{r^2-t^2}$, then as $r\to\infty$ the main cylinder becomes a plane, $(r-\sqrt{r^2-t^2})\to 0$, and (6.14) becomes $q(t)=p(t)\cos\beta$, as expected.

(e) When $\cos\beta\neq 0$, (6.14) can be solved for $p(t)$ to give

$$p(t)=q(t)\sec\beta-m(t)\tan\beta, \tag{6.15}$$

a linear combination of the two profile functions $q(t)$ and $m(t)$. In particular, if the cutting cylinder is a circular cylinder of radius a cutting a main circular cylinder of radius r at angle β, then (6.15) holds with $q(t)=\sqrt{a^2-t^2}$ and $m(t)=r-\sqrt{r^2-t^2}$.

Example 11 (Intersection of two circular cylinders). Now take the special case of (e) in which $a=r$. Let $z=p(t)$, and write (6.15) in the form

$$z+r\tan\beta=\sqrt{r^2-t^2}(\sec\beta+\tan\beta)$$

or

$$\frac{t^2}{r^2}+\left(\frac{z+r\tan\beta}{r(\sec\beta+\tan\beta)}\right)^2=1.$$

This is the equation of an ellipse with center at $(0,-r\tan\beta)$ in the tz plane and semiaxes of lengths r and $r(\sec\beta+\tan\beta)$. Because the projected curve C_p is an ellipse, we know that the unwrapped curve C_u will be sinusoidal.

Example 12 (Tilted view of a geodesic). Example 2 revealed that a geodesic on a right circular cylinder is a circular helix whose side view is a sine curve. On a circular cylinder of radius 1, the geodesic with unwrapping function $u(x)=cx$ has horizontal profile $p(t)=c\arcsin t$, and the main cylinder has circular profile $m(t)=1-\sqrt{1-t^2}$. Hence (6.14) gives

$$q(t)=c(\arcsin t)\cos\beta+(1-\sqrt{1-t^2})\sin\beta.$$

When $\tan\beta = c$, the helix is viewed along one of its tangents of constant slope, and this becomes

$$\frac{1}{\sin\beta}q(t) = \arcsin t + (1 - \sqrt{1-t^2}).$$

The right member is easily shown to represent a cycloid, the path traced out by a point on the circumference of a circular disk that rolls along a straight line, so the profile $q(t)$ describes a cycloid dilated in the z' direction by the factor $\sin\beta$.

This can also be demonstrated physically with a flexible spring, such as the toy called "slinky." By stretching the spring and viewing it from different directions you can see the helix change its appearance from sine curve to curtate cycloid, to cycloid, and to prolate cycloid.

6.10 UNWRAPPING CURVES FROM A GENERAL CYLINDER

The examples treated thus far involve curves unwrapped from a right circular cylinder. Now we discuss a curve unwrapped from a noncircular vertical cylinder. For the directrix we take a logarithmic spiral because its arclength is easily calculated.

Example 13 (Logarithmic spiral cylinder cut by a plane). A plane curve with polar equation $r = e^{c\theta}$, where c is a positive constant and θ varies over all real values, is commonly called a *logarithmic spiral*. Here $c = \tan\delta$, where δ is the complement of the constant angle between the tangent to the spiral and the radial line from the origin, as shown in Figure 6.18a.

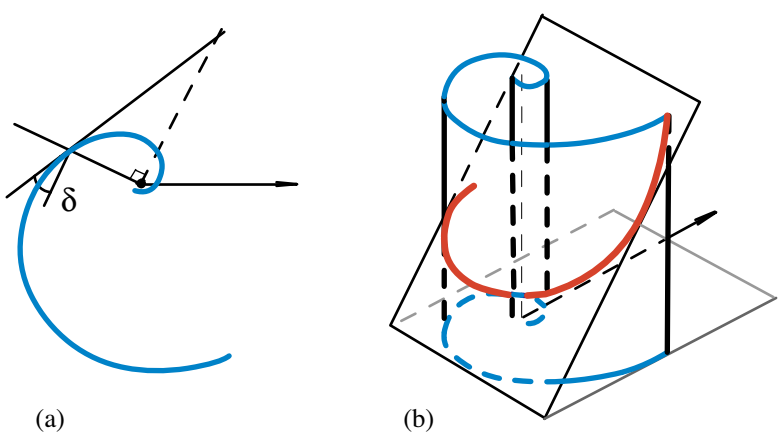


Figure 6.18: (a) Logarithmic spiral. (b) Vertical cylinder with the spiral as directrix cut by an inclined plane.

The graph makes infinitely many circuits around the origin, and Figure 6.18a shows part of the graph when $c = 1$ and θ varies over a finite interval. Figure 6.18b shows a vertical cylinder having this portion of the spiral as directrix. A plane inclined at 45° cuts the cylinder along a curve, part of which is shown Figure 6.18b,

and we wish to find a cartesian equation of the form $z = u(x)$ for the unwrapped curve when the cylinder is unwrapped onto the xz plane.

To do this, we first determine the arclength of the spiral $r = e^{c\theta}$. The integral for arclength in polar coordinates is easily calculated and shows that as θ varies over any finite interval $[a, b]$ the corresponding arclength of the spiral is $k(e^{cb} - e^{ca})$, where

$$k = \frac{\sqrt{1+c^2}}{c}.$$

Geometrically, $k = 1/\cos\delta$. In particular, when $a \to -\infty$, the arclength s of the spiral from the origin to a point at radial distance r is given by $s = kr$. The horizontal projection t of the point (r, θ) is

$$t = r\cos\theta = \frac{s}{k}\cos(\log\frac{s}{k}),$$

which expresses t as a function of s. If $x = s(t)$, then $t = s^{-1}(x)$, where

$$s^{-1}(x) = \frac{x}{k}\cos(\log\frac{x}{k}).$$

A plane through the origin inclined at 45° with the t axis has cartesian equation $z = p(t) = t$. Therefore by Theorem 6.1 the unwrapping function is $u(x) = p(s^{-1}(x))$, so the unwrapping equation becomes

$$z = \frac{x}{k}\cos(\log\frac{x}{k}).$$

Figure 6.19 shows the general shape of the graph of the unwrapped curve, greatly distorted horizontally. The slopes of the dotted lines will change if the angle of inclination of the cutting plane changes.

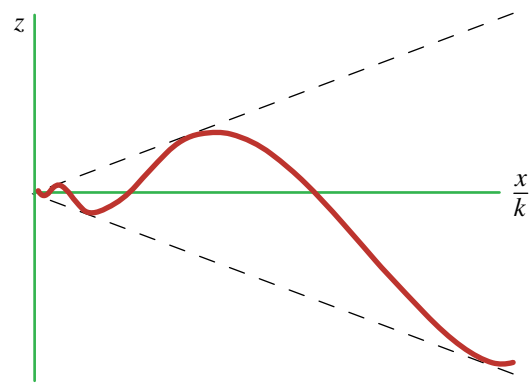


Figure 6.19: Unwrapped curve of intersection of a vertical logarithmic cylinder cut with an inclined plane.

6.11 APPLICATIONS TO GRAPHICS

Powerful 3-D modeling programs can be used to render the qualitative shape of the curve of intersection of two cylindrical surfaces on a computer screen. However, exact equations like those derived in this paper provide a deeper understanding, and are also useful when graphing projection and unwrapping functions. The graphs are not specified by 3-D modeling programs which, for example, do not reveal whether a projected oval curve is an ellipse or a curve of higher degree. Knowing that a curve is an ellipse can have profound implications. For example, Kepler's landmark discovery that planetary orbits are elliptic implies Newton's inverse-square law of gravitation.

The exact equations allow us to easily plot curves and animate them on a computer screen using simple 2-D graphics programs instead of 3-D programs. Most illustrations in this chapter were prepared in this manner.

Remarks on arclength invariance.

We have seen that arclength invariance plays a fundamental role in determining the shape of a curve unwrapped from a cylinder. Although arclength calculations usually involve complicated integrals, we know that distances are preserved when a developable surface such as a cylinder is unwrapped. Thus, a curve on a cylinder has the same arclength as its unwrapped counterpart. In particular, from Figure 6.1 we see that an ellipse has the same arclength as an unwrapped sine curve, without the need to calculate the elliptic integrals that produce numerical values.

In the rest of this chapter, we replace the main cylinder by a right circular cone, and investigate what happens to a curve on its lateral surface when the cone is unwrapped onto a plane. The analysis on a cone differs from that on a cylinder, but again depends on arclength invariance, and leads to interesting and somewhat surprising results expressed as simple formulas.

PART 2: UNWRAPPING FROM CONES

6.12 UNWRAPPING CURVES FROM A RIGHT CIRCULAR CONE

Start with a curve C lying on the surface of a right circular cone, and consider questions of the following type:

What is the shape of the image of C when the cone is unwrapped, that is, tipped on a generator and rolled onto a plane, or onto another cone? How does C appear when viewed from different directions?

All cones in this chapter are right circular cones, and we unwrap not only conic sections, but any curve lying on the surface of a cone. We employ three simple geometric transformations: projecting the curve onto the *ceiling plane* (a plane orthogonal to the cone's axis and passing through its vertex), scaling the ceiling projection radially from the vertex, and compressing the polar coordinate angle.

We formulate the basic questions more precisely in terms of equations, and show they can be answered once the ceiling projection is known. We discuss interesting curves on cones that are not conic sections, including geodesic curves, and curves cut by various cylinders. These reveal some surprising results. For example, when a hyperbolic cylinder cuts a cone, the curve of intersection may appear in different directions not only as a hyperbola, but also as an ellipse, a parabola, or even a geodesic (on a special cone).

Figure 6.20a shows two familiar curves on a cone. One is a circular cross section we call a *base*, whose unwrapped image is a circular arc, the dashed curve in Figure 6.20b. The other is an ellipse that unwraps to form a new plane curve, a *generalized ellipse* shown in Figure 6.20b.

Now we replace the ellipse in Figure 6.20a by a general curve C lying on a cone, and unwrap it onto a plane. In the plane, generators of the cone are mapped onto radial lines emanating from the cone's vertex. A point P on C is mapped onto a point in the plane with polar coordinates $(R(\theta), \theta)$, with the origin at the vertex of the cone, where $R(\theta)$ is the distance of P from the vertex of the cone, and θ is the

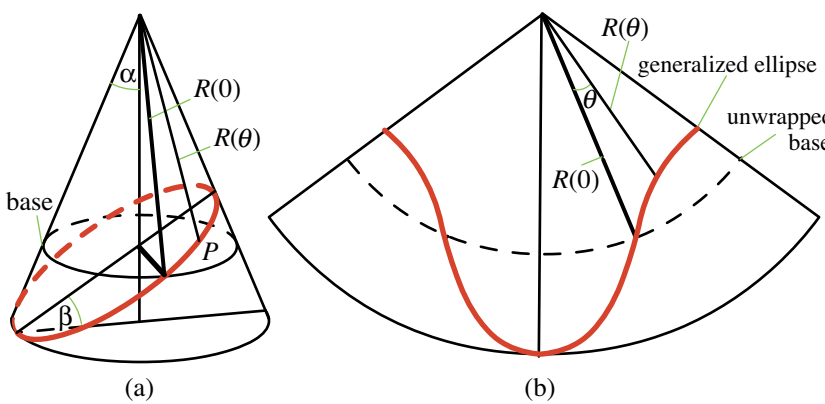


Figure 6.20: (a) An elliptical cross section. (b) Its unwrapped image on a plane.

polar angle in radians measured from a fixed radial line to that through the image of P. We can also regard θ as being measured along the surface of the cone from the fixed generator to the generator through P. Thus, $(R(\theta), \theta)$ can be thought of as conical coordinates on the cone itself. The function $R(\theta)$ depends on C, and we formulate the following general problem:

Basic problem: *For a curve C on the cone, obtain an explicit formula for $R(\theta)$. In particular, describe $R(\theta)$ when C is a conic section.*

The analysis on a cone differs from that on a cylinder, but again relies on arclength invariance, as indicated in the following special case.

6.13 UNWRAPPED BASE AND PRESERVATION OF ARCLENGTH

For a finite portion of the cone with a circular base, as shown in Figure 6.21a, the unwrapped image of the base is a circular arc with center at the cone's vertex and radius equal to the slant height s of the finite portion. In this simple case, the basic problem is easily solved because the radial distance $R(\theta)$ is constant, $R(\theta) = R(0) = s$.

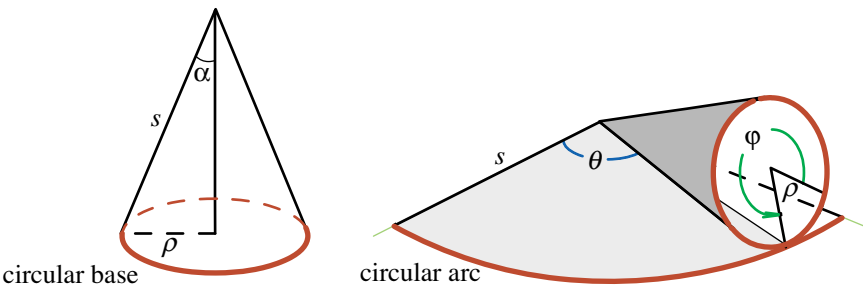


Figure 6.21: (a) Finite portion of a cone with slant height s. (b) Unwrapping the surface of the finite portion.

Figure 6.21 also reveals a basic fact that plays a key role in solving the general problem. Let ρ denote the radius of the base in Figure 6.21a. When the base rolls through an angle of φ radians, the corresponding portion of the base of arclength $\rho\varphi$ unwraps onto a circular arc of radius s and central angle that we denote by θ (Figure 6.21b). Because the cone is a developable surface, distances are preserved when the cone is unrolled onto a plane, so we have

$$s\theta = \rho\varphi. \tag{6.16}$$

It is easy to see that the relation between θ and φ is independent of ρ and s. In Figure 6.21a, α is half the vertex angle of the cone ($0 < \alpha < \pi/2$), and ρ is related to s by

$$\rho = s \sin\alpha. \tag{6.17}$$

Combining (6.16) and (6.17) we find a relation independent of ρ and s :

$$\theta = \varphi \sin\alpha. \tag{6.18}$$

The simple relation (6.18) occurs repeatedly in analyzing the shape of a curve unwrapped from a cone. With $k = 1/\sin\alpha$, (6.18) can be written as

$$\varphi = k\theta. \tag{6.19}$$

Thus, the sine of half the vertex angle determines the relation between φ and θ.

When the cone is cut by a plane through a diameter of the base inclined at an angle β, the conic section is an ellipse, parabola, or hyperbola, depending on β. Figure 6.20a shows an ellipse, which unwraps to form a generalized ellipse that oscillates about the image of the base as shown in Figure 6.20b. An explicit formula for $R(\theta)$, which depends on both β and the cone's vertex angle, is given in (6.25).

6.14 REFORMULATED PROBLEM IN TERMS OF CEILING PROJECTION

In Part 1 of this chapter we analyzed a general curve on a cylinder by projecting it onto an unwrapping plane parallel to the generators of the cylinder. To analyze a curve C lying on a cone, we project it upward onto the horizontal ceiling plane, a plane orthogonal to the axis of the cone and passing through its vertex V, as indicated in Figure 6.22.

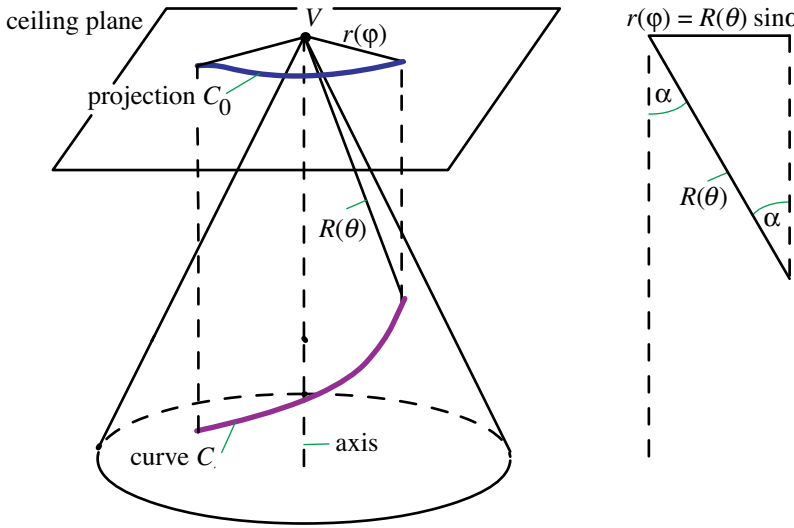


Figure 6.22: Curve C on the cone projects onto curve C_0 in the ceiling plane with polar equation $r = r(\varphi)$.

The curve C projects onto a curve C_0 in the ceiling plane that we describe with a polar equation $r = r(\varphi)$, where $r(\varphi)$ is the radial distance measured from V as origin. The *ceiling projection* C_0 is the profile of a vertical cylinder that intersects the cone along C. Figure 6.22 reveals that the distances $R(\theta)$ and $r(\varphi)$ satisfy $r(\varphi) = R(\theta) \sin\ \alpha$, where α is half the vertex angle of the cone. This simple relation, together with (6.19), shows that the ceiling projection is the key that unlocks the basic problem, as revealed in Theorem 6.6.

Theorem 6.6. *Let C be a curve on the surface of a cone with vertex angle 2α. If the ceiling projection C_0 has polar equation $r = r(\varphi)$, then the unwrapped image of C has polar equation*

$$R(\theta) = kr(k\theta), \tag{6.20}$$

where $k = 1/\sin\alpha$. Conversely, if $R(\theta)$ is known, then (6.20) determines $r(\varphi)$:

$$r(\varphi) = R(\varphi/k)/k. \tag{6.21}$$

Proof. The relation $r(\varphi) = R(\theta)\sin\alpha$ becomes $r(\varphi) = R(\theta)/k$ which, in view of (6.19), gives (6.20) and (6.21).

Before discovering Theorem 6.6, we solved the basic problem for the special case of an unwrapped conic, using a lengthy brute force analysis of the solid geometry of the cone. To our surprise, the resulting formula for $R(\theta)$ in (6.25) resembled that of an ordinary conic. An analysis of this formula suggested introducing the ceiling projection, which applies not only to conic sections, but to any curve C lying on a cone.

6.15 CEILING PROJECTION AND UMBRELLA TRANSFORMATION

Equation (6.21) is the end result of three transformations: projecting C onto C_0, which produces $r(\varphi)$; stretching each radial distance $r(\varphi)$ by the factor k; and squeezing the polar angle φ by the factor $1/k$. The first two can be combined into one transformation given by

$$R(\theta) \;=\; kr(\varphi), \tag{6.22}$$

which is a scaled version of the ceiling projection.

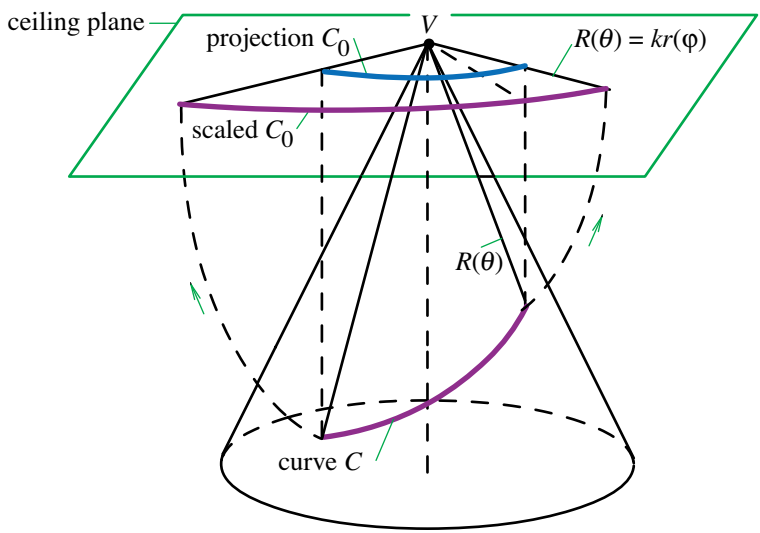


Figure 6.23: An umbrella transformation (6.22) maps C from the cone onto a scaled version of projection C_0.

To visualize (6.22) geometrically, regard the cone as an umbrella that can be opened up flat onto the ceiling plane by rotating each generator vertically upward about V, as suggested by the example in Figure 6.23. The rotation preserves radial distances from the vertex, but increases angles between generators on the cone. Two generators separated by an angle θ measured along the surface of the cone lie on two vertical planes through the axis making a dihedral angle φ, and during the

rotation the umbrella transformation stretches the angle between them from θ to $\varphi = k\theta$.

6.16 CONE TO CONE

The analysis used to prove Theorem 6.6 also treats the more general case in which a curve C_1 on one right circular cone with vertex angle $2\alpha_1$ is unwrapped onto a curve C_2 on another right circular cone having the same vertex but with vertex angle $2\alpha_2$. When the vertex angle $2\alpha_2$ is a straight angle, the second cone becomes an unwrapping plane. When we unwrap one cone onto another it is understood that we keep the cones tangent to each other along a common rolling generator.

The next theorem relates the ceiling projection functions of curves C_1 and C_2.

Theorem 6.7. *If the ceiling projection of curve C_1 has polar equation $r_1 = r_1(\varphi)$, and that of curve C_2 has polar equation $r_2 = r_2(\varphi)$, then*

$$r_2(\varphi) = \mu r_1(\mu\varphi), \tag{6.23}$$

where $\mu = \sin\alpha_2 / \sin\alpha_1$.

Proof. When we unwrap one cone with vertex angle $2\alpha_1$ onto another having the same vertex but with vertex angle $2\alpha_2$, angle θ and distance $R(\theta)$ from the common vertex to a common point P on the two curves are the same on both cones. In other words, both curves obviously coincide when unwrapped onto a plane. Hence by (6.20) we have $R(\theta) = k_1 r_1(k_1\theta)$ and $R(\theta) = k_2 r_2(k_2\theta)$, which implies

$$r_2(k_2\theta) = (k_1/k_2) r_1(k_1\theta).$$

Let $\mu = k_1/k_2 = \sin\alpha_2/\sin\alpha_1$, and let $\varphi = k_2\theta$. Then $k_1\theta = \mu\varphi$ and the foregoing equation becomes (6.23).

When $k_2 = 1$, cone 2 coincides with its ceiling plane, $\varphi = \theta$, $r_2 = R$, $r_1 = r$, and (6.23) turns into (6.20). Note also that $\mu < 1$ if cone 1 has a larger vertex angle than cone 2, and, vice versa, $\mu > 1$ if cone 2 has a larger vertex angle than cone 1.

6.17 UNWRAPPING A CONIC SECTION FROM A CONE TO A PLANE

Now take C to be a conic section cut from a cone by a plane inclined at an angle β with the ceiling plane, where $0 \le \beta < \pi/2$. The cutting plane intersects the axis of the cone at a point O that we use as the center of a circular base of radius ρ (Figure 6.24). As before, α is half the vertex angle of the cone, where $0 < \alpha < \pi/2$. In Figure 6.24, C is shown as an ellipse, but the analysis also applies to a parabola or hyperbola. The generator through point P on C intersects the base at point B whose polar coordinates in the plane of the base are (ρ, φ), where φ is measured from OA, where $A = (\rho, 0)$. The case $\beta = 0$ corresponds to unwrapping the base, which was treated earlier. Figure 6.24 also shows the ceiling projection C_0 of C. The following theorem solves the unwrapping problem for a conic, and also includes a surprising result in part (a).

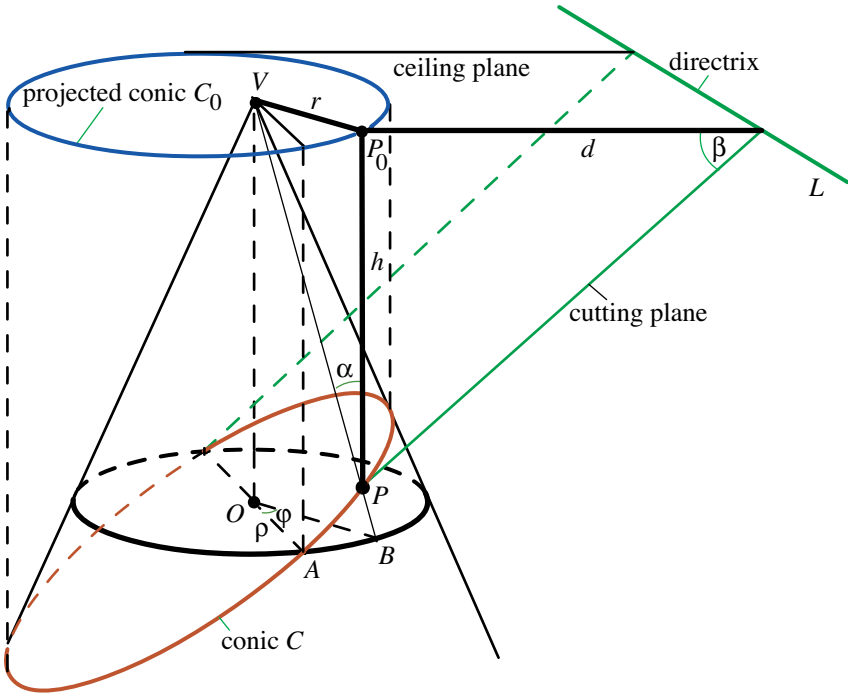


Figure 6.24: Diagram for proving parts (a), (b), (c) of Theorem 6.8. The ceiling plane intersects the cutting plane along the directrix of the projected conic.

Theorem 6.8. *The ceiling projection of a conic section C is another conic C_0 with*
(a) *a focus at the vertex of the cone,*
(b) *a directrix at the line of intersection of the cutting plane and the ceiling plane,*
(c) *eccentricity $\lambda = \tan\alpha \tan\beta$,*
(d) *polar equation*

$$r(\varphi) = \frac{r(0)}{1 + \lambda \sin\varphi}. \tag{6.24}$$

Thus, if $k = 1/\sin\alpha$, the unwrapped image of C on a plane has polar equation

$$R(\theta) = \frac{kr(0)}{1 + \lambda \sin(k\theta)}. \tag{6.25}$$

Proof. Let L denote the line of intersection of the cutting plane and ceiling plane. For a point P on C, let P_0 denote its projection on C_0. Let d be the distance from P_0 to L, and r the distance from P_0 to V, as shown in Figure 6.24.

We now show that the eccentricity ratio r/d is $\tan\alpha \tan\beta$, a constant (independent of P_0) that we denote by λ. This will prove (a), (b), and (c). Write the ratio r/d as

$$\frac{r}{d} = \frac{r}{h}\,\frac{h}{d},$$

where h is the distance from P to P_0. From Figure 6.24 we infer that $r/h = \tan\alpha$ and $h/d = \tan\beta$, hence $r/d = \tan\alpha\tan\beta$, as required.

To derive (6.25), use the focus V as origin in the ceiling plane, and let $r(\varphi)$ denote the distance from V to the point on C_0 with polar coordinates $(r(\varphi), \varphi)$, where φ is measured from a line through the focus parallel to the directrix, as in Figure 6.25.

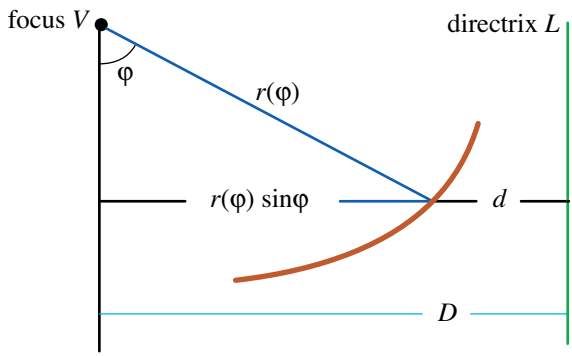


Figure 6.25: Diagram for deriving the polar equation of projected conic C_0.

Because C_0 is a conic with eccentricity λ, the focal definition of conic gives $r(\varphi) = \lambda d$. But $d = D - r(\varphi)\sin\varphi$, where D is the distance from the focus to the directrix, so $r(\varphi) = \lambda(D - r(\varphi)\sin\varphi)$, which when solved for $r(\varphi)$ gives $r(\varphi) = \lambda D/(1 + \lambda\sin\varphi)$. When $\varphi = 0$ we get $\lambda D = r(0)$, which proves (6.24), and (6.25) follows by Theorem 6.6.

The ratio $e = (\sin\beta)/(\cos\alpha)$ is known to be the eccentricity of the conic section C in Figure 6.24. It is easily verified that $e = \lambda = 1$ when $\alpha + \beta = \pi/2$, that $0 < e < 1$ and $0 < \lambda < 1$ when $\alpha + \beta < \pi/2$, and that $e > 1$ and $\lambda > 1$ when $\alpha + \beta > \pi/2$. Therefore C and its ceiling projection C_0 are of the same type: ellipse, parabola, or hyperbola. Although their eccentricities may differ, both are simultaneously less than 1, equal to 1, or larger than 1.

In Theorem 6.8, the relations between the parameters λ, k and the angles α, β imply the restrictions $\lambda \geq 0$, $k \geq 1$. The inequality $\lambda \geq 0$ is not serious, because changing the sign of λ in (6.25) is equivalent to replacing θ by $-\theta$, which means the unwrapping occurs in the opposite direction. The restriction $k \geq 1$ is more serious because $k = 1/\sin\alpha$. However, (6.25) is meaningful for all real λ and k and gives a function $R(\theta)$, periodic in θ with period $2\pi/k$, that represents a well-defined curve even if $k < 1$. This motivates the following notion of a generalized conic.

Definition. For constants $R_0 \geq 0$, $\lambda \geq 0$, and real k, a plane curve described by the polar equation

$$R(\theta) = \frac{R_0}{1 + \lambda\sin(k\theta)}, \tag{6.26}$$

is called a *generalized conic*.

The curve is called a generalized ellipse, parabola, or hyperbola, according as $\lambda < 1$, $\lambda = 1$, or $\lambda > 1$, respectively.

If $k = 1$, the cone is its ceiling plane, and (6.26) is the polar equation in this plane of a conic with eccentricity λ and with one focus at the origin. If $k > 1$, Theorem 6.8 tells us that the curve in (6.26) is obtained by unwrapping a conic section of eccentricity λ from a cone with vertex at the origin and vertex angle 2α, where $\sin\alpha = 1/k$. If $k < 1$, the curve in (6.26) cannot be obtained by unwrapping a conic section from a cone onto a plane, but it can be realized as the ceiling projection of a curve C on a cone K' with vertex angle $2\alpha'$, where $\sin\alpha' = k$, and where C is obtained by wrapping a conic of eccentricity λ from a plane onto K'. In terms of equations, a conic with polar equation (6.24) is the unwrapped version of a curve C on a cone K' with vertex at a focus of (6.24) and with ceiling projection described by (6.25).

6.18 EXAMPLES OF GENERALIZED CONICS

Examples are shown in Figures 6.26 through 6.31. In Figures 6.26, 6.27, 6.28, the angle θ runs through one period interval of length $2\pi/k$, and in Figure 6.29 through more than one period interval. In Figures 6.26 to 6.29, each example has $k > 1$ and can be obtained by unwrapping a conic section from a cone onto a plane. In Figures 6.30 and 6.31, each example has $k < 1$, namely $k = 1/2$, and cannot be obtained by unwrapping a conic from a cone.

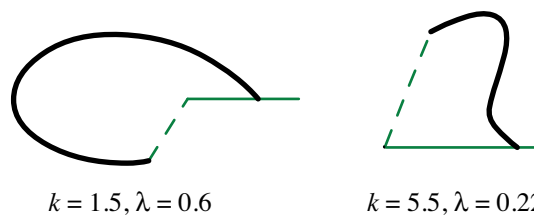


Figure 6.26: Generalized ellipses, one period.

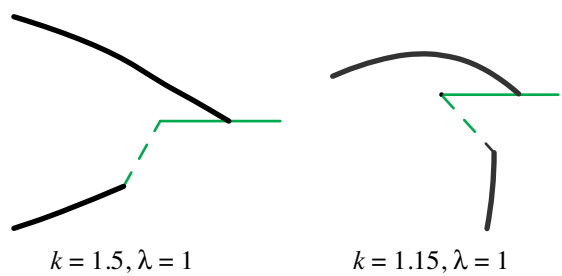


Figure 6.27: Generalized parabolas, one period.

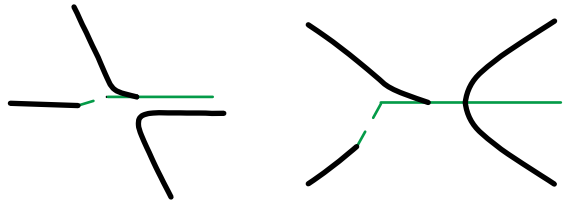


Figure 6.28: Generalized hyperbolas, one period. Both nappes are cut and unwrapped.

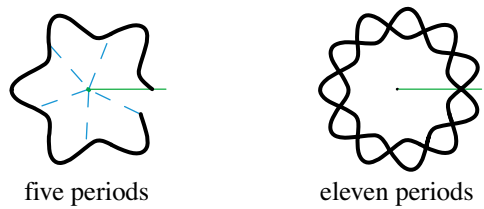


Figure 6.29: Generalized ellipses as in Figure 6.26b ($k = 5.5$, $\lambda = 0.22$), with more than one period.

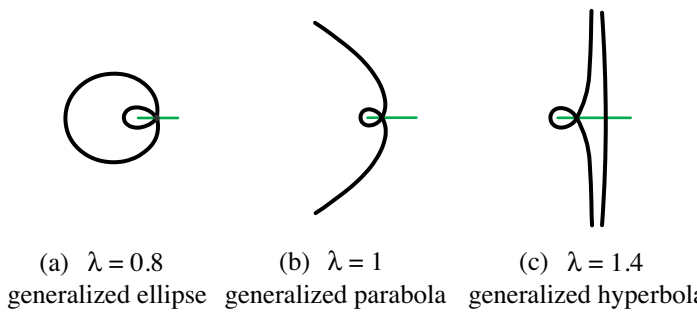


Figure 6.30: Generalized conics with $k = 0.5$ in (6.26). They cannot be obtained by unwrapping a conic from a cone, but each is the ceiling projection of a conic wrapped from a plane onto a cone.

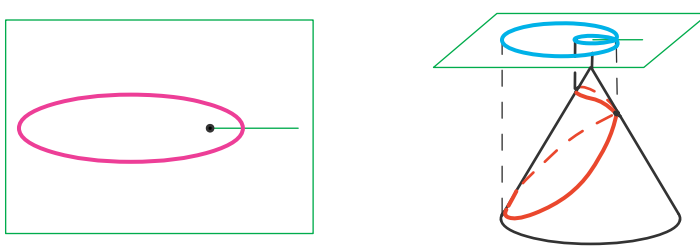


Figure 6.31: Generalized ellipse in Fig. 6.30a as ceiling projection of a curve obtained by wrapping an ellipse onto a cone with vertex angle $\pi/3$ to form a curve with two loops.

6.19 LIMITING CASES

From a given cone, one can obtain all possible ellipses and parabolas as conic sections, but not all hyperbolas, only those whose asymptotes intersect at an angle smaller than the vertex angle of the cone. By contrast, all possible conics can be obtained as limiting cases of (6.26), as will be shown presently.

First we show that sinusoidal curves unwrapped from circular cylinders are limiting cases of generalized conics. In the plane of the unwrapped cone (Figure 6.20b), the difference $y = R(0) - R(\theta)$ is the radial distance from the image of the circular base to the generalized conic. From (6.26) we have $R_0 = R(0)$, and we obtain

$$y = R(0)\frac{\lambda \sin(k\theta)}{1 + \lambda \sin(k\theta)}. \tag{6.27}$$

Keep the radius ρ of the base fixed and keep the angle of inclination β fixed, but let $\alpha \to 0$ so the vertex of the cone recedes to infinity. Then the cone becomes a cylinder of radius ρ (which can be regarded as the limiting case of a cone).

What happens to the right-hand side of (6.27) as $\alpha \to 0$?

For small α, we can approximate $\tan\,\alpha$ by $\sin\alpha$, so the product $\lambda = \tan\alpha\tan\beta$ can be approximated by $\sin\alpha\tan\beta$. The denominator of (6.27) is very close to 1, so its right member has the approximate value

$$R(0)\sin\alpha\tan\beta\,\sin\varphi,$$

where $\varphi = k\theta$. But, in view of (6.17), $R(0)\sin\alpha = \rho$, hence (6.27) is nearly the same (for small α) as the limiting relation

$$y = \rho\,\tan\beta\sin\varphi = \rho\,\tan\beta\sin(x/\rho),$$

where x is the length of arc subtended by an angle φ on a circle of radius ρ. This is a cartesian equation of a sinusoidal curve cut from a circular cylinder of radius ρ by a plane inclined at angle β. In other words, if the circular base of the cone is kept fixed while the vertex recedes to infinity, the cone becomes a cylinder, and the generalized ellipse unwrapped from the cut cylinder becomes a sinusoidal curve, as introduced in Section 6.2.

The other limiting case is when $\alpha \to \pi/2$ and the cone flattens onto the ceiling plane. In this case we keep $R(0)$ and λ fixed, so that $\tan\beta = \lambda/\tan\alpha$. Then $\tan\alpha \to \infty$, $\beta \to 0$, and $k = 1/\sin\alpha \to 1$, and the limiting value of the polar equation (6.25) is

$$R(\theta) = \frac{R(0)}{1 + \lambda\,\sin\,\theta}. \tag{6.28}$$

This describes an ordinary conic section of eccentricity λ, where $R(\theta)$ is the distance from a focus to the point $(R(\theta), \theta)$ on the conic, as shown in Figure 6.25. Geometrically, as $\alpha \to \pi/2$ and the cone flattens onto its ceiling plane, with $\tan\beta = \lambda/\tan\alpha$, the conic section of eccentricity λ turns into the conic described by (6.28), which is the ceiling conic (6.24). In other words, as the cone flattens onto a plane,

the limiting case of the conic section coincides with the ceiling conic having its focus at the vertex, and it also coincides with the limiting case of the generalized conic. All three types of conics, ellipse, parabola and hyperbola, with all possible values of the eccentricity, can occur in the limiting case.

6.20 OTHER CURVES ON A CONE

In addition to conic sections, there are other interesting curves on a cone that we can analyze using our transformations. In each of the next two examples, the curve is determined by specifying its ceiling projection C_0 to be a spiral, and we find that the unwrapped curve is a spiral of the same type.

Example 14 (Archimedean spiral). Here $r(\varphi) = c\varphi$ for a constant $c > 0$. In the ceiling plane, C_0 intersects a radial line at equidistant points, with distance $2\pi c$ between consecutive intersections. The curve C spirals around the cone as shown in Figure 6.32. From (6.20) we find

$$R(\theta) = ck^2\theta.$$

The unwrapped curve is an Archimedean spiral with c replaced by ck^2.

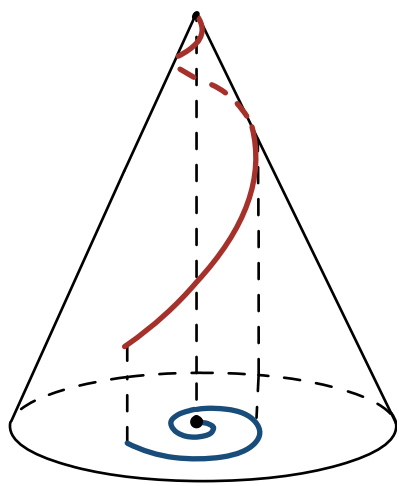


Figure 6.32: A conical spiral, with an Archimedean spiral as ceiling projection and as unwrapped curve.

Example 15 (Logarithmic spiral). Now $r(\varphi) = Ae^{c\varphi}$ for constants $A > 0$, $c > 0$. In the ceiling plane, the tangent line to the spiral at each point makes a constant angle δ with the radial line to that point, where $\cot\delta = c$. From (6.20) we obtain $R(\theta) = kAe^{ck\theta}$, hence the unwrapped curve is another logarithmic spiral with new constants. Its tangent line makes a constant angle ψ with the radial line, where $\cot\psi = ck$.

Example 16 (Geodesic on a cone). Because distances are preserved when a cone is unwrapped, the image of a geodesic arc on a cone (the shortest path joining two points on the surface) is a line segment in the plane of the unwrapped cone. To construct a geodesic curve on a cone, start with a straight line and wrap it onto the cone. Figure 6.33a shows a line L and a point V not on the line that we take as the vertex of an unwrapped cone. The entire line has polar equation

$$R(\theta) = \frac{d}{\cos\theta},$$

where d is the shortest distance from V to the line, and θ varies from $-\pi/2$ to $\pi/2$. The plane determined by the line and V can be rolled into many right circular cones with V as a common vertex but with different vertex angles, and the line is mapped onto a geodesic curve on each such cone. If the cone has vertex angle 2α, the ceiling projection of the geodesic has polar equation

$$r(\varphi) = \frac{d/k}{\cos(\varphi/k)},$$

where $k = 1/\sin\alpha$ and $-k\pi/2 < \varphi < k\pi/2$. Figures 6.33b and c show one example of a geodesic and its ceiling projection. Figure 6.34 shows more ceiling projections.

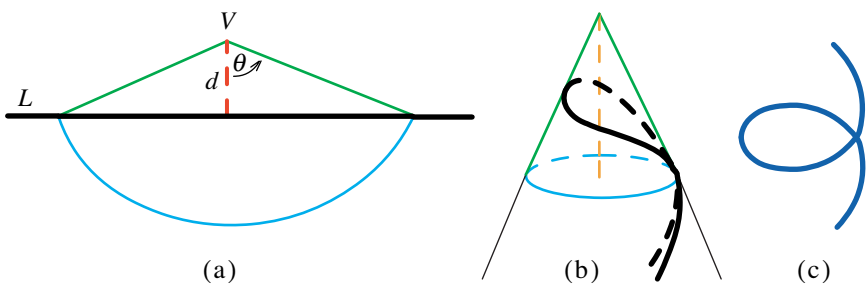


Figure 6.33: Line segment in (a) wrapped onto a geodesic on a cone in (b), with ceiling projection (c).

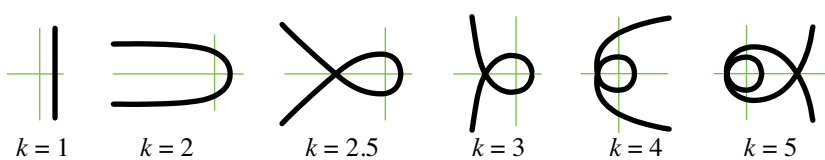


Figure 6.34: Ceiling projections of one line wrapped onto cones with different vertex angles ($\sin\alpha = 1/k$).

6.21 VERTICAL WALL PROJECTION

Analyzing the unwrapped version of a curve on the lateral surface of a cone is equivalent, by Theorem 6.6, to finding its ceiling projection. We turn now to examples in which the curve C is the intersection of the cone with a horizontal cutting cylinder whose generators are parallel to the ceiling plane. The projection of C on a vertical plane perpendicular to the generators is a profile of the cylinder.

We choose such a plane through the axis of the cone and call it the *wall plane*. It intersects the ceiling plane along a line we designate as the t axis, with its origin at V, as shown in Figure 6.35. The axis of the cone is designated the z axis, but with its positive direction pointing down as indicated in Figure 6.35. If the cone is flipped over and the ceiling plane becomes a horizontal floor plane, the coordinate axes of the wall plane will be in traditional position, with the positive t axis pointing to the right, and the positive z axis pointing up.

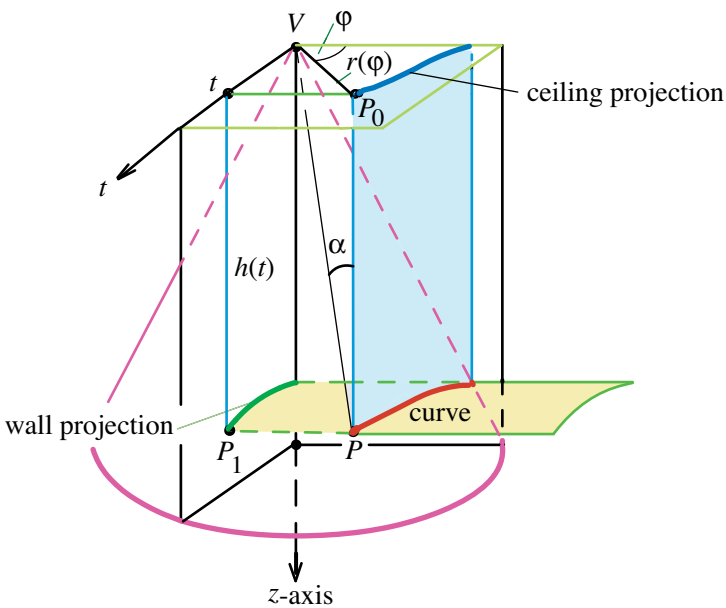


Figure 6.35: Relating the wall projection and ceiling projection of a curve on a right circular cone.

The cutting cylinder intersects the wall plane along a profile curve with implicit equation of the form $p(t, z) = 0$. We call this curve the *wall projection* of C. A point P on C has point P_0 as ceiling projection with polar coordinates (r, φ), where φ is measured as indicated in Figure 6.35, and $r = r(\varphi)$ is the ceiling function. The point P also has wall projection P_1 with coordinates (t, z) related by $p(t, z) = 0$. The next theorem, which follows at once from Figure 6.35, relates the coordinates (t, z) of P_1 with the polar coordinates (r, φ) of P_0.

Theorem 6.9. *On a right circular cone with vertex angle 2α, let $c = \tan\,\alpha$, and let C be a curve with ceiling projection function $r = r(\varphi)$, and wall profile $p(t, z) = 0$. Then the coordinates are related by*

$$t = r\sin\varphi,\ \ z = \frac{r}{c}. \tag{6.29}$$

Consequently,

$$\varphi = \arcsin(\frac{t}{r}),\ \ r = cz, \tag{6.30}$$

and

$$p(r\,\sin\varphi, r/c) = 0.$$

In particular, if the wall profile gives z explicitly as a function of t, say $z = h(\,t), \text{then}$

$$r(\varphi) = ch(r(\varphi)\,\sin\varphi) \tag{6.31}$$

and

$$ch(t) = r(\arcsin(\frac{t}{ch(t)})). \tag{6.32}$$

The following examples specify the wall projection and determine $r(\varphi)$.

Example 17 (Linear wall projection: $h(t) = at + b$). The cutting cylinder in this case is a plane, and (6.31) becomes

$$r(\varphi) = c(ar(\varphi)\sin\varphi + \,b),$$

which can be solved for $r(\varphi)$ to yield

$$r(\varphi) = \frac{bc}{1 - ac\sin\,\varphi}.$$

As expected, this is the polar equation of a conic section.

Example 18 (Circular cutting cylinder of radius 1). Cut the cone with a circular drill of radius 1, perpendicular to the axis of the cone, whose center in the tz plane is at the point (w, d). Then the wall projection satisfies the implicit equation

$$(t - w)^2 + (z - d)^2 = 1.$$

From (6.29) we find $(r(\varphi)\sin\varphi - w)^2 + (\frac{r(\varphi)}{c} - d)^2 = 1$, which is quadratic in the ceiling projection function $r(\varphi)$.

If $w = 0, \text{the axis of the cutting cylinder passes through the axis of the cone and}$ the quadratic equation simplifies to

$$(r(\varphi)\sin\varphi)^2 + \,(\frac{r(\varphi)}{c} - d)^2 = 1. \tag{6.33}$$

A graphing calculator can draw the graphs of (6.33), revealing the ceiling projection of the hole for various values of c and d. Figure 6.36a shows snapshots on

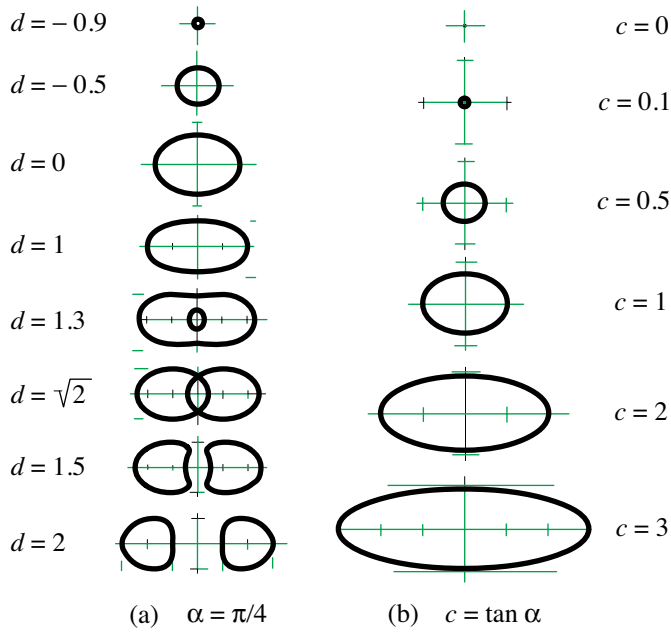


Figure 6.36: Snapshots of ceiling projection of a horizontal hole of radius 1 drilled through the axis. In (a) the cone is fixed and the coordinate d varies. In (b) the drill is fixed at $d = 0$ and the vertex angle changes.

one nappe only for $c = 1$ and increasing d, and Figure 6.36b shows snapshots for $d = 0$ and increasing c.

Powerful 3-D modeling programs can be used to render the qualitative shape of the curve of intersection of a cone and a cutting cylinder. But exact equations like those derived here provide a deeper understanding, and are also useful when graphing projection functions such as those in Figure 6.36 with simple 2-D programs.

In Figure 6.36a the vertex angle 2α of the cone is $\pi/2$. For small d the projection is an oval curve, which gradually changes its size and shape as d increases. At some stage it is pierced by a hole that increases in size until $d = \sqrt{2}$, when the ceiling projection turns into two overlapping confocal ellipses. As d increases further, the ceiling projection splits into two symmetrically disconnected pieces that move further apart.

In Figure 6.36b, the axis of the drill passes through the vertex of the cone and the snapshots show how the ceiling projection of the hole varies as the vertex angle of the cone increases. We were surprised to learn that all the projected curves in Figure 6.36b are ellipses! This is easily verified by writing (6.33) in rectangular coordinates. Incidentally, the example with $d = 0$ in Figure 6.36a is one of the ellipses.

6.22 CEILING AND WALL PROJECTIONS OF A ROTATED CURVE

To view a curve on a cone from different directions perpendicular to the axis of the cone, we simply rotate the cone about its axis to see how the ceiling projection changes, then use Theorem 6.9 to determine the corresponding wall projection. Specifically, take a curve C on a cone with ceiling projection $r = r(\varphi)$, and rotate the cone through an angle γ about its axis, measured counterclockwise when viewed from above the ceiling plane. The rotation replaces φ by $\varphi + \gamma$, so the ceiling projection $r_\gamma(\varphi)$ of the rotated curve is

$$r_\gamma(\varphi) = r(\varphi + \gamma). \tag{6.34}$$

As the next example shows, rotation can alter dramatically the appearance of the wall projection.

Example 19 (Parabolic cutting cylinder). Cut the cone with a parabolic cylinder whose wall projection is $z = b - at^2$, where a and b are constants, $a \neq 0$. The wall projection is a parabola that opens downward if $a > 0$ and upward if $a < 0$ (because the z axis points down.) By (6.29) the corresponding ceiling projection satisfies

$$\frac{r(\varphi)}{c} = b - ar^2(\varphi)\sin^2\varphi.$$

When expressed in rectangular coordinates, with

$$x = r(\varphi)\cos\varphi, \quad y = r(\varphi)\sin\varphi,$$

this equation has degree 4 in y and degree 2 in x.

Now rotate the cone through a right angle ($\gamma = \pi/2$). By (6.34), the rotated curve has ceiling projection $r_{\pi/2}(\varphi)/c = b - ar^2_{\pi/2}(\varphi)\cos^2\varphi$, which can also be written as

$$\frac{r_{\pi/2}(\varphi)}{c} = b - ar^2_{\pi/2}(\varphi) + ar^2_{\pi/2}(\varphi)\sin^2\varphi. \tag{6.35}$$

To find the wall projection of the rotated curve, replace $r_{\pi/2}(\varphi)$ by cz and replace $r_{\pi/2}(\varphi)\sin\varphi$ by t, using (6.29). Then (6.35) becomes

$$t^2 - c^2z^2 - \frac{z}{a} = \frac{b}{a}. \tag{6.36}$$

This represents a hyperbola in the tz plane, regardless of the sign of a. In other words, when a horizontal parabolic cylinder intersects a cone along a curve C, the horizontal parabolic profile of this curve, when rotated through a right angle, becomes a hyperbola! When the parabolic cylinder is tangent to two generators of the cone, the hyperbola is degenerate, and the ceiling projection is two intersecting confocal parabolas. The limiting case of (6.36) with $c = 0$ is worth noting. As $c \to 0$, the vertex angle of the cone tends to 0, the cone becomes a right circular cylinder, and (6.36) becomes $z = at^2 - b$. This is another derivation of a result found in Section 6.6 (Figure 6.10). When a vertical circular cylinder is cut by a parabolic cylinder and rotated by a right angle, the parabolic profile is flipped over.

When the same analysis is applied to the parabolic cutting cylinder whose wall projection is $t = b - az^2$ the ceiling projection satisfies the equation $r(\varphi)\sin\varphi = b - ar^2(\varphi)/c^2$, which, in rectangular coordinates, has the form

$$x^2 + (y + \frac{c^2}{2a})^2 = \frac{c^2}{4a^2}(c^2 + 4ab).$$

When $c^2 + 4ab > 0$ this represents a circle. In other words, a parabolic cylinder with profile symmetric about the z axis cuts each nappe of the cone along a circle. This also shows that if the cone is drilled by a circular cylinder with axis parallel to the axis of the cone, the edge of the hole will appear as a parabola when viewed on the vertical wall projection. If the cone is rotated through a right angle, the wall projection of the rotated curve is given by $c^2z^2 - t^2 = (b - az^2)^2$, which has degree 4 in z and degree 2 in t.

Example 20 (Central conic as profile of cutting cylinder; ceiling projection). Cut the cone with a horizontal cylinder whose profile is a central conic (ellipse or hyperbola). A central ellipse is given by

$$(\frac{t}{a})^2 + (\frac{z}{b})^2 = 1, \tag{6.37}$$

whereas a hyperbola is given by one of

$$(\frac{t}{a})^2 - (\frac{z}{b})^2 = 1, \tag{6.38}$$

or

$$(\frac{z}{b})^2 - (\frac{t}{a})^2 = 1, \tag{6.39}$$

depending on whether the axis through the foci is horizontal or vertical. The elliptic cylinder and the second hyperbolic cylinder in (6.39) always cut the cone, but the first hyperbolic cylinder in (6.38) may or may not, depending on how the angle between its asymptotes compares with the vertex angle of the cone.

The elliptic case is simplest so we treat it first. In this case (6.37) and (6.29) give us

$$r^2(\frac{b^2c^2}{a^2}\sin^2\varphi + 1) = b^2c^2. \tag{6.40}$$

In rectangular coordinates with origin at the cone's vertex we have $\sin\varphi = y/r$, and (6.40) can be written as $a^2r^2 + b^2c^2y^2 = a^2b^2c^2$, which, on replacing r^2 by $x^2 + y^2$, becomes

$$a^2x^2 + (a^2 + b^2c^2)y^2 = a^2b^2c^2.$$

This shows that the ceiling projection is always an ellipse with its center at the vertex.

The same argument, applied to (6.38), gives

$$(b^2c^2 - a^2)y^2 - a^2x^2 = a^2b^2c^2. \tag{6.41}$$

This represents a hyperbola if $b^2c^2 > a^2$ and the empty set if $b^2c^2 < a^2$. This can also be seen geometrically. Recall that $c = \tan\alpha$, and note that the slopes of the asymptotes are $\pm a/b$. When $b^2c^2 < a^2$ the angle between the asymptotes is greater than the vertex angle of the cone and the hyperbolic cylinder does not cut the cone. But if $b^2c^2 > a^2$ the hyperbolic cylinder cuts the cone into two separate pieces that project onto the ceiling plane as two branches of the hyperbola in (6.41). If $b^2c^2 = a^2$ the hyperbolic cylinder degenerates to two planes whose angle of intersection is the vertex angle of the cone, and the ceiling projection consists of a single point, the vertex.

When the same argument is applied to (6.39) we find

$$a^2x^2 + (a^2 - b^2c^2)y^2 = a^2b^2c^2. \tag{6.42}$$

In this case, the ceiling projection is an ellipse or a hyperbola, according as $b^2c^2 < a^2$ or $b^2c^2 > a^2$. If $b^2c^2 < a^2$ the hyperbolic cylinder cuts the cone along a curve that appears as a hyperbola when viewed along the horizontal direction of the cylinder, but as an ellipse when viewed along the cone's axis.

Example 21 (Central conic as profile of cutting cylinder; rotated wall projection). Now rotate the cone in Example 20 through a right angle and determine the wall projection of the rotated curve. First we determine the ceiling projection of the rotated curve.

When we apply (6.34) with $\gamma = \pi/2$ to (6.40), which comes from the elliptic cutting cylinder in (6.37), the term $\sin^2\varphi$ is replaced by $\cos^2\varphi = 1 - \sin^2\varphi$, and r becomes $r_{\pi/2}$. In the rotated ceiling polar equation, replace $r_{\pi/2}$ by cz and $\sin\varphi$ by $t/(cz)$, to obtain

$$(a^2 + b^2c^2)z^2 - b^2t^2 = a^2b^2, \tag{6.43}$$

which represents a hyperbola. In other words, an elliptic cutting cylinder intersects a cone along a curve C that appears as an ellipse (expected) when viewed along the horizontal direction of the cylinder, but as a hyperbola (unexpected!) when viewed from a perpendicular horizontal direction.

We can do the same for each of the hyperbolic cutting cylinders in (6.38) and (6.39). From (6.41) we find the rotated wall projection is given by

$$(b^2c^2 - a^2)z^2 - b^2t^2 = a^2b^2. \tag{6.44}$$

In this case the cylinder cuts the cone only if $b^2c^2 > a^2$, in which case (6.44) represents a hyperbola. In other words, both the ceiling projection (6.41) and the rotated wall projection are hyperbolas.

The situation is different for the second hyperbolic cutting cylinder in (6.39). The ceiling projection in (6.42) is an ellipse if $b^2c^2 < a^2$ and a hyperbola if $b^2c^2 > a^2$. When the cone is rotated through a right angle, the rotated wall projection is given by $(a^2 - b^2c^2)z^2 + b^2t^2 = a^2b^2$.

This is an ellipse if $b^2c^2 < a^2$ and a hyperbola if $b^2c^2 > a^2$, so the rotated wall projection is the same type of curve as the ceiling projection.

Example 22 (Side view of a particular geodesic). In Example 16 we found that a geodesic on a cone has ceiling projection function

$$r(\varphi) = \frac{d/k}{\cos(\varphi/k)}.$$

A surprising result occurs on a cone with $k = 2$ (vertex angle $2\alpha = \pi/3$). The ceiling projection satisfies

$$r^2(\varphi) = 2d^2/(1 + \cos\,\varphi).$$

When rotated through a right angle, this gives a new ceiling projection function $r^2(\varphi) = 2d^2/(1-\sin\varphi)$,with corresponding wall projection

$$cz(cz - t) = 2d^2,$$

where $c = \tan(\pi/6)$.

The quadratic form $cz^2 - czt$ has a negative discriminant, so the wall projection is a hyperbola. In other words, on this cone, a side view of a geodesic appears as exactly half of one branch of a hyperbola.

Examples 19 through 22 reveal some of the unexpected results referred to in Section 6.12. If a hyperbolic cylinder cuts a cone, the curve of intersection appears in different directions not only as a hyperbola, but also as an ellipse, a parabola, or even a geodesic.

6.23 TILTED WALL PROJECTION

In the foregoing discussion, the cutting cylinder has generators parallel to the ceiling plane. In descriptive geometry, and in applications to sheet metal work, the cutting cylinder is not always parallel to the ceiling plane but may be tilted at an angle β. By tilting the wall plane about the t axis we can relate the profile of C on the tilted plane with the original wall and ceiling projections.

Introduce a tilted z' axis by rotating the vertical z axis about the horizontal t axis by an angle β, as shown in Figure 6.37. Let $z' = q(t)$ denote the explicit equation of the profile of the tilted cylinder, and let P be a point on the curve C of intersection of the tilted cylinder with the cone. Point P has wall projection P_1 and ceiling projection P_0. Figure 6.37 shows P, P_0, and P_1 in a vertical plane perpendicular to the t axis. In this plane, viewed along the t axis, $q(t)$ is the sum of two lengths, $h(t)\cos\,\beta$ and $x\,\sin\,\beta$, where $x = r(\varphi)\cos\,\varphi$ and $h(t)$ is the distance from P_0 to P. Therefore we have the following equation relating the tilted profile function $q(t)$ with the wall profile $h(t)$ and the ceiling projection $r(\varphi)$:

$$q(t) = h(t)\cos\beta + r(\varphi)\cos\varphi\,\sin\beta. \tag{6.45}$$

The wall projection function $h(t)$ is related to the ceiling projection $r(\varphi)$ by Theorem 6.9, and (6.45) determines the tilted profile function $q(t)$ in terms of the

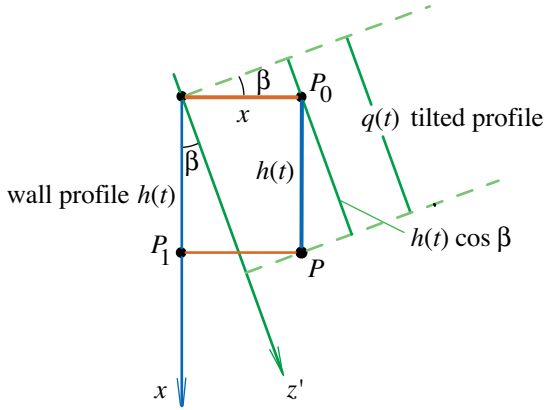


Figure 6.37: Profile of a tilted cutting cylinder expressed in terms of the wall profile and ceiling projection.

tilting angle β. To relate $q(t)$ and $h(t)$ directly, use Theorem 6.9 to get $r(\varphi) = ch(t)$, and rewrite (6.45) as

$$q(t) = h(t)\cos\beta + ch(t)\cos\varphi\sin\beta. \tag{6.46}$$

For a relation not involving φ, recall from (6.29) that $ch(t)\sin\varphi = t$. Hence

$$ch(t)\,\cos\varphi = ch(t)\sqrt{1-\sin^2\varphi} = \sqrt{c^2h^2(t)-t^2}$$

and (6.46) takes the form (6.47) in the following theorem.

Theorem 6.10. *A tilted cylinder whose generators make an angle β with the ceiling plane, and which has tilted profile function $z' = q(t)$, intersects a cone of vertex angle 2α along a curve whose vertical wall profile $z = h(t)$ is related to $q(t)$ by*

$$q(t) = h(t)\cos\beta + \sqrt{c^2h^2(t)-t^2}\,\sin\beta, \tag{6.47}$$

where $c = \tan\alpha$.

This expresses the tilted profile $q(t)$ directly in terms of the wall profile $h(t)$ and the angles α and β. Also, if $q(t)$ is given, we can use (6.47) to find $h(t)$ by solving a quadratic equation. For example, if $q(t) = at + b$, the slanted cutting cylinder is a plane. If we put $z = h(t)$, then (6.47) becomes quadratic in z and t, and the horizontal profile is a conic, as expected.

6.24 ARCLENGTH AND AREA

When a curve of length L on a cone is unwrapped onto another curve, the arclength L remains unchanged because distances are preserved. For example, the generalized ellipse in Figure 6.20b has the same length as the ellipse in Figure 6.20a, even though

there is no simple formula for calculating the arclengths. In general, any unwrapped curve has the same length as its wrapped version on the cone.

Unwrapping also preserves areas. In Figure 6.21, a portion of the circular cone unwraps onto a circular sector with central angle θ as shown in Figure 6.21b, with area $s^2\theta/2$. This sector, the unwrapped portion of the cone, has the same area, which, by (6.17), is given by $\rho s\varphi/2$. In particular, when $\varphi = 2\pi$, each of the areas is equal to $\pi\rho s$.

More generally, we can ask for the sectorial area $A(\theta_1, \theta_2)$ of the region bounded by the unwrapped image of any curve C on the cone and two rays $\theta = \theta_1$ and $\theta = \theta_2$ emanating from the origin. When $\theta_1 < \theta_2$ the area is given by

$$A(\theta_1, \theta_2) = \frac{1}{2}\int_{\theta_1}^{\theta_2} R^2(\theta)\, d\theta, \tag{6.48}$$

where $R(\theta)$ is determined by (6.26). Because areas are preserved when unwrapping a cone, the integral has the same value as the lateral surface area of the portion of the cone between the vertex and the original curve C on the cone. The change of variable $\varphi = k\theta$ transforms the integral in (6.48) to

$$A(\theta_1, \theta_2) = \frac{1}{2k}\int_{k\theta_1}^{k\theta_2} R^2(\varphi/k)\, d\varphi.$$

Because of (6.21) we can write this as

$$A(\theta_1, \theta_2) = k(\frac{1}{2}\int_{\varphi_1}^{\varphi_2} r^2(\varphi)\, d\varphi) \tag{6.49}$$

where $r(\varphi)$ is the ceiling projection function for C_0, and $\varphi_i = k\theta_i$. The factor multiplying k in (6.49) is the sectorial area of the region in the ceiling plane bounded by C_0 and the rays $\varphi = \varphi_1$ and $\varphi = \varphi_2$. Equation (6.49) can be stated as

Theorem 6.11. *On a cone of vertex angle 2α, the lateral surface area of the portion of the cone between the vertex and an arc of the original curve C on the cone is k times the area of the corresponding ceiling projection, where $k = 1/\sin\alpha$.*

This is to be expected, because a thin triangle of area T formed by two nearby generators on the lateral surface of the cone, with one vertex at the vertex of the cone, projects onto the ceiling plane onto a triangle with area $T_0 = T\sin\alpha$, hence $T = kT_0$.

Although Theorem 6.11 refers to an unwrapped sectorial region, it implies a more general result for regions lying between two curves C_1 and C_2 with corresponding unwrapping functions R_1 and R_2, and two rays $\theta = \theta_1$ and $\theta = \theta_2$. The integral

$$\frac{1}{2}\int_{\theta_1}^{\theta_2} |R_2^2(\theta) - R_1^2(\theta)|\, d\theta$$

gives the area of both the unwrapped region and the corresponding region on the cone. Each of the areas is k times the area of the corresponding ceiling projection.

Again this is to be expected because an elemental region of area T on the cone projects onto the ceiling plane as a region of area $T_0 = T \sin\alpha$, so $T = kT_0$. The next example gives a surprising consequence of this result.

Example 23 (Surface area cut from a cone by a vertical drill). In Example 19 we learned that a circular drill with its axis parallel to that of a cone cuts a hole on the surface of the cone whose edge appears (unexpectedly) as a parabola when viewed on the vertical wall projection. The geometric argument supporting Theorem 6.11 gives another unexpected result. The region on the surface of the cone has area k times that of the ceiling projection, which is constant for a vertical drill, even if the hole is not circular. In other words, the surface area removed from a cone by a vertical puncturing tool of cross-sectional area A is equal to kA, regardless of the shape or location of the tool.

We conclude with an interesting observation concerning the elliptical cross section cut by a plane inclined at angle β as shown Figure 6.20a. Let E denote the area of the elliptical disk, and let S denote the lateral surface area of the finite portion of the cone with vertex angle 2α cut off by the disk. On the one hand, the ceiling projection of the ellipse has area $E\cos\beta$, and on the other hand, it is $S\sin\alpha$. This simple result for the ellipse must surely be known, but we could not find it in the literature. It deserves to be better known, so we state it here as a theorem.

Theorem 6.12. *The area E of an elliptic cross-sectional disk, and the lateral surface area S of the finite portion of the cone cut off by the plane of the disk, satisfy*

$$S\sin\alpha = E\cos\beta.$$

NOTES ON CHAPTER 6

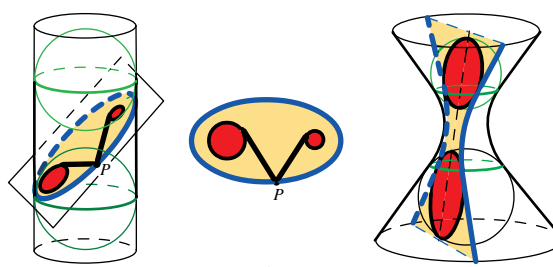
Most of the material in this chapter was originally published in [19], which was awarded a Lester R. Ford Award in 2008.

Chapter 7

NEW DESCRIPTIONS OF CONICS VIA TWISTED CYLINDERS, FOCAL DISKS, AND DIRECTORS

These problems can be easily solved by the methods developed in this chapter. The reader may wish to try solving them before reading the chapter.

An ellipse is a cross section of a circular cylinder cut by an inclined plane. Two spheres inscribed in the cylinder intersect the cutting plane to form two circular disks as in (a), called focal disks, shown in (b) in the cutting plane.



Prove that the sum of the lengths of the tangent segments from P to the two focal disks is constant.

For an ellipse, find two focal disks such that the absolute difference of the lengths of the tangent segments from a point P on the ellipse to the disks is constant.

The two shaded disks in (c) are plane sections of spheres inscribed in a hyperboloid of revolution. *What can be said about the sum and difference of lengths of the tangent segments from a point on the hyperbolic cross section to the two disks*?

CONTENTS

From any point of an ellipse the sum of distances to its two foci is constant. For a hyperbola the absolute difference of these distances is constant. These properties are generalized to provide a unified characterization of all conics, including the parabola! We inscribe Dandelin-type spheres in a twisted cylinder (hyperboloid) rather than a cone, and pierce the cutting plane to produce focal disks. Focal distances are replaced by lengths of tangent segments to these disks. For each conic and any pair of focal disks, the sum of tangent distances is a constant on some portions of the conic, while on the remaining portions their difference is the same constant! Each conic now has infinitely many focal disks, resulting in a rich variety of configurations, some of which cannot occur on a cone. Special cases reveal many surprises. For example, the absolute difference of tangent distances can be constant everywhere on an ellipse or a parabola! Our approach also leads naturally to a generalization of the classical focus-directrix-eccentricity description of conics.

7.1 INTRODUCTION

Conics have been investigated since ancient times as sections of a circular cone. Surprising descriptions of these curves are revealed by investigating them as sections of a hyperboloid of revolution, referred to here as a *twisted cylinder*. We generalize the classical focus-directrix property of conics by what we call the *focal disk-director property* (Section 7.2). We also generalize the classical bifocal properties of central conics by the *bifocal disk property* (Section 7.5), which applies to all conics, including

the parabola. Our main result (Theorem 7.5) is that the two generalized properties are satisfied by sections of a twisted cylinder, and by no other curves.

Twisted cylinders.

A circular cylinder is a ruled surface with its generators parallel to the axis of the cylinder. Figure 7.1a shows a portion of a cylinder between two circular bases. Rotate the lower circle about the axis to form a new ruled surface (Figure 7.1b),

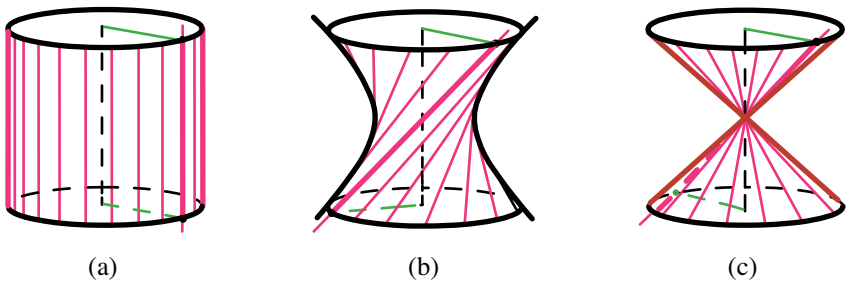


Figure 7.1: Cylinder (a) and cone (c) as special cases of twisted cylinder (b).

all of whose generators make the same angle with the axis. Because the constancy of this angle is fundamental in our analysis, we prefer to call the surface a twisted cylinder rather than a hyperboloid of revolution. The circular cylinder and cone are special cases. (For interactive Java animation see the website cited at the end of this chapter.) In this chapter, all twisted cylinders have vertical axes.

In general, a section of a twisted cylinder by an inclined plane is analogous to a section of a cone (Figure 7.2). For a small angle of inclination, the section is an ellipse. When the cutting plane is tilted more to become parallel to a generator, the section is a parabola. Tilting it further produces a hyperbola.

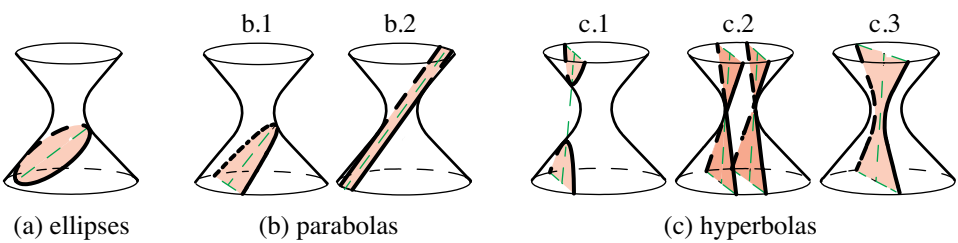


Figure 7.2: Intersections of a twisted cylinder with an inclined plane.

Differences between sections of a cone and of a twisted cylinder.

Significant differences are revealed when the cutting plane is translated. On a cone, translation of the cutting plane always produces a similar conic, with the same eccentricity. But on a twisted cylinder, as Figure 7.2b shows, a parabola b.1 can degenerate into a pair of parallel lines b.2 (which we call a *degenerate parabola*).

More dramatic changes occur when the intersection is a hyperbola, c.1. Figure 7.2c shows two critical positions, c.2, at which the plane is tangent to the twisted cylinder, and the hyperbola degenerates into a pair of intersecting lines. Between these critical positions there are intermediate flipped hyperbolas, c.3, whose eccentricity is not the same as that of the hyperbolas in c.1. Further translation flips the hyperbola again, to sections similar to those in c.1. On a cone, the critical positions coincide, and the flipped hyperbolas c.3 do not appear. In this chapter, the term *conic* refers to any section of a twisted cylinder, including degenerate cases.

Focal disks.

Inscribe a sphere inside a twisted cylinder so it intersects the cutting plane along a circular disk, shown shaded in Figure 7.3. This disk is in the plane of the conic, and we call it a *focal disk* for that conic. If the sphere happens to be tangent to the

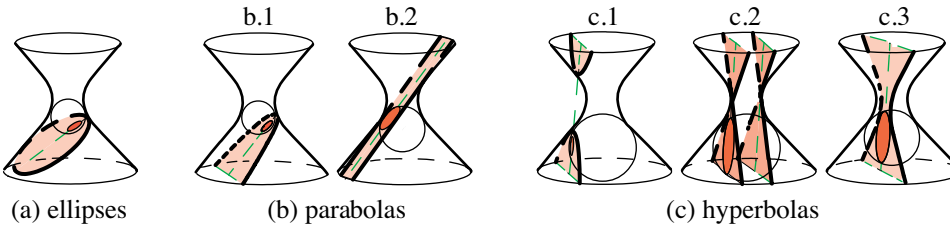


Figure 7.3: Inscribed sphere pierces the plane of the conic along a focal disk.

cutting plane, the focal disk is a point (which turns out to be a focus!). This can occur in Figure 7.3 in (a), b.1, and c.1 but not in b.2 and c.3. In c.3 the foci of the hyperbola are outside the twisted cylinder.

Families of focal disks.

By moving the sphere upward or downward through the cutting plane, keeping it inscribed in the twisted cylinder, we obtain an infinite family of focal disks for a given conic. Examples in the plane of the conic are shown for an ellipse in Figure

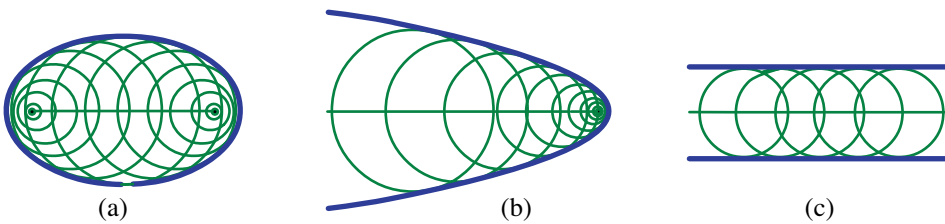


Figure 7.4: Families of focal disks associated with the conics in Figures 7.3a,b.

7.4a, a parabola in Figure 7.4b, and a degenerate parabola in Figure 7.4c. When the ellipse is a circle the focal disks are concentric with it.

Figures 7.5a, b, and c show the focal disks obtained by moving an inscribed sphere in Figure 7.3c, corresponding to the sections in c.1, c.2, and c.3.

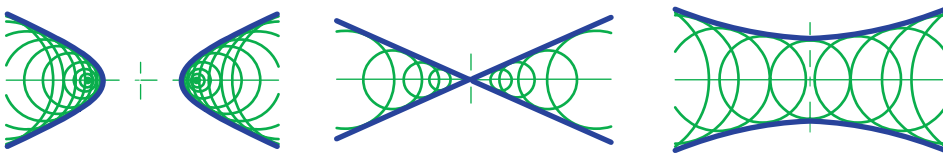


Figure 7.5: Families of focal disks associated with the sections in Figure 7.3c. The family in (c) does not contain the foci of the hyperbola and cannot occur on a cone.

Abnormal configurations.

A conic with its focal disks defines a configuration. Those configurations in Figure 7.4c and Figure 7.5c we call *abnormal* to emphasize that they cannot occur on a cone. In them a focal disk cannot be shrunk to a focus, and a circular disk tangent to both branches is a focal disk. The degenerate configuration in Figure 7.4c can be regarded as a limiting configuration of Figure 7.5c. The configurations that can occur on a cone we call *normal*.

The rest of the chapter is divided into three parts as follows:

Part 1: Focal disk-director description of noncircular conics (Sections 7.2-7.4).
Part 2: Bifocal disk description of conics (Sections 7.5-7.7).
Part 3: Supplementary results (Sections 7.8-7.13).

PART 1: FOCAL DISK-DIRECTOR DESCRIPTION OF NONCIRCULAR CONICS

7.2 DISK-DIRECTOR RATIO. FOCAL DISK-DIRECTOR PROPERTY

In Figure 7.6a the inscribed sphere touches the twisted cylinder along a circle we call a *terminator*, terminology borrowed from astronomy. If the cutting plane is parallel to the plane of the terminator, the section is a circle and the focal disk is concentric with the circle. Otherwise, the planes intersect along a line we call a *director*. The section shown in Figure 7.6 is an ellipse, but it could be any noncircular conic.

In Figure 7.6b, let P denote a point on the conic, let PT be the length of a tangent segment from P to the focal disk, and let PD be the distance from P to the director.

We introduce the *disk-director ratio* $q = PT/PD$ and show that the conic has the following *focal disk-director property*:

Proposition 1. *On a noncircular conic of intersection, the disk-director ratio* $q = PT/PD$ *is constant, that is, independent of* P.

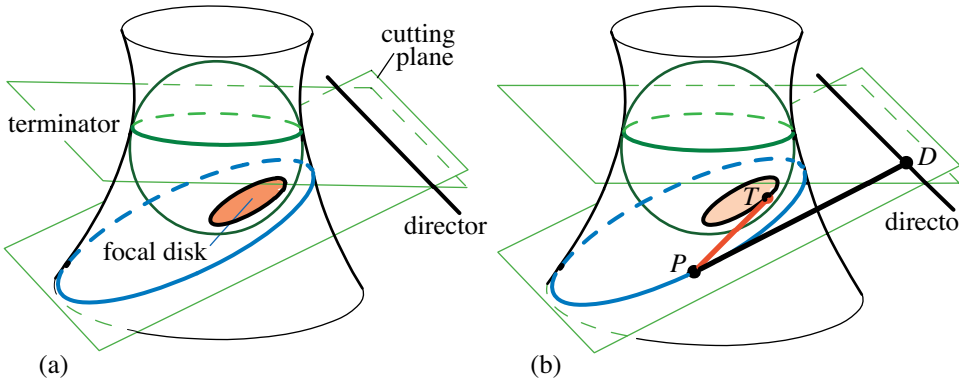


Figure 7.6: (a) Focal disk and director. (b) Disk-director ratio $q = PT/PD$ is constant.

Proof. In Figure 7.7, α denotes the angle formed by a generator and a line parallel to the axis of the twisted cylinder, and $\beta > 0$ denotes the angle between the cutting plane and a plane through the terminator. Let P_1 denote the point where the generator through P intersects the terminator, and let P_0 denote the vertical projection of P on the plane of the terminator. Then $PP_0/PP_1 = \cos\,\alpha$, and $PP_0/PD = \sin\,\beta$. But $PT = PP_1$ because both are lengths of tangent segments to the same sphere from an external point P. Therefore

$$q = \frac{PT}{PD} = \frac{PP_1}{PD} = \frac{PP_1}{PP_0}\frac{PP_0}{PD} = \frac{\sin\,\beta}{\cos\,\alpha}, \tag{7.1}$$

which is a constant independent of P. The case of a flipped hyperbola is shown in Figure 7.7b.

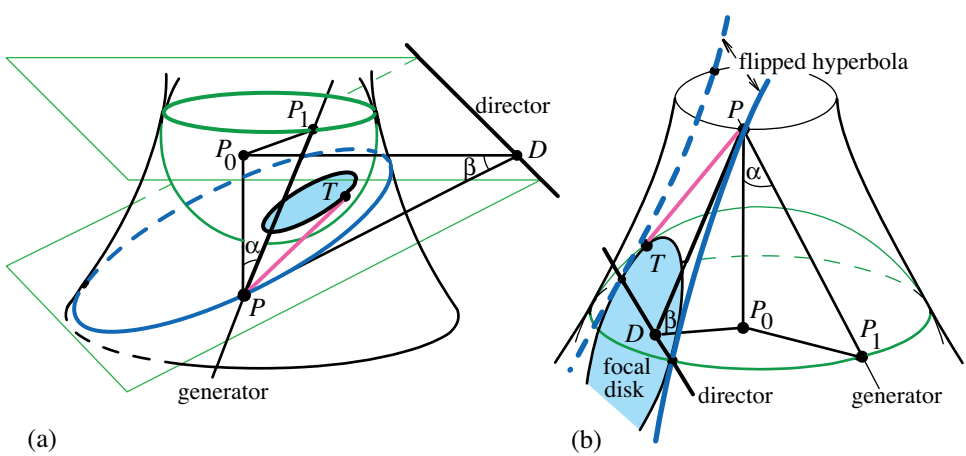


Figure 7.7: Proof of Proposition 1: (a) ellipse; (b) flipped hyperbola.

If $\beta = 0$ the cross section is a circle, and the definition of director D as a line of

intersection does not apply. But in this case we can regard D as a line at infinity and the ratio $q = PT/PD$ as being zero. In most results involving the disk-director ratio we assume $q > 0$ and assume that the conic is noncircular without explicitly saying so.

Invariant properties of q.

From (7.1) we see that the disk-director ratio q depends only on β, the angle of inclination of the cutting plane, and on α, the angle between a generator of the twisted cylinder and its vertical axis. Therefore q is invariant under any change in configuration that preserves the angles, such as moving the inscribed sphere upward or downward, translating the cutting plane, or scaling by similarity. When the twisted cylinder is a cone with vertex angle 2α, (7.1) is a known formula for the eccentricity of the conic section. By moving the inscribed sphere through the cutting plane to produce families of focal disks, we obtain:

Proposition 2. *A conic of intersection has the focal disk-director property with respect to infinitely many focal disk-director pairs, all of which have the same disk-director ratio.*

7.3 DISK-DIRECTOR RATIO RELATED TO ECCENTRICITY

The classical focus-directrix property (Figure 7.8b) does not apply to circular conic sections, and likewise the generalized focal disk-director property does not apply to circular sections of a twisted cylinder. For a nonhorizontal cutting plane, the focal disk-director property in this plane is illustrated in Figure 7.8a for normal configurations. It states that there is a constant q such that $PT = q\,PD$ for every P on the conic. In normal cases, when the focal disk can shrink to a point (denoted by F in Figure 7.8b), the director becomes the classical directrix of a conic with focus F. The ratio $q = PT/PD$, which is unchanged in the shrinking process, turns into the classical eccentricity $e = PF/PD$ for a noncircular conic section. To preserve the relation $q = e$ for all conics we take $q = 0$ for a circle. Therefore:

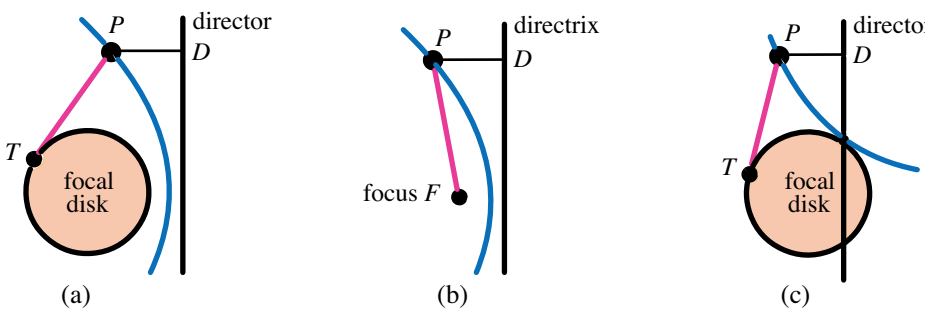


Figure 7.8: Focal disk-director property: $PT = q\,PD$ (a) normal configuration; (c) abnormal configuration. (b) Classical focus-directrix property : $PF = e\,PD$.

For normal configurations, the disk-director ratio is equal to the eccentricity of the conic.

In an abnormal case shown in Figure 7.5c and in Figure 7.8c, the disk cannot be shrunk to a focus, and the foregoing argument does not apply. The transition from normal to abnormal occurs when the hyperbola c.1 in Figure 7.3c degenerates and then flips to the hyperbola c.3 which, as we will see in a moment, has similar asymptotes but not necessarily the same eccentricity. During the transition, q remains invariant and is equal to the eccentricity e of the hyperbola in c.1.

To determine the relation between q and the eccentricity ε of the flipped hyperbola with the same asymptotes it suffices to relate e and ε. We do this with the help of Figure 7.9.

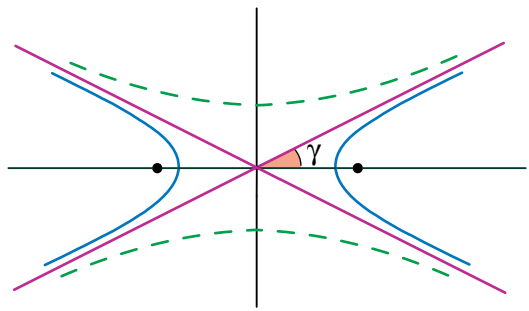


Figure 7.9: The hyperbola with horizontal focal axis has eccentricity $e = 1/\cos\,\gamma$. The flipped hyperbola with same asymptotes has conjugate eccentricity $\varepsilon = 1/\sin\,\gamma$.

Conjugate eccentricities of flipped hyperbolas with the same asymptotes.

In Figure 7.9, γ denotes the angle between an asymptote of a hyperbola of eccentricity e and its horizontal focal axis. It is easy to show that $e = 1/\cos\gamma$. (See Figure 13.14 in [1].) Hence $q = 1/\cos\gamma$, so the angle between asymptotes is also invariant under translation of the cutting plane. But the eccentricity ε of the flipped hyperbola (dashed in Figure 7.9) with the same asymptotes is given by $\varepsilon = 1/\cos(\pi/2 - \gamma) = 1/\sin\gamma$. Because $\cos^2\gamma + \sin^2\gamma = 1$ we find

$$\frac{1}{e^2} + \frac{1}{\varepsilon^2} = 1, \text{ or } e = \frac{\varepsilon}{\sqrt{\varepsilon^2 - 1}}.$$

We call e and ε *conjugate eccentricities*. They are equal only for a rectangular hyperbola, in which case the asymptotes are perpendicular and $e = \varepsilon = \sqrt{2}$.

Relation between the disk-director ratio and eccentricity.

We know that $q = e$ for the configurations in Figures 7.5a and 7.5c, where e is the eccentricity of the hyperbola in Figure 7.5a. Consequently, for the flipped hyperbola

in Figure 7.5c, q is related to its eccentricity ε by

$$q = \frac{\varepsilon}{\sqrt{\varepsilon^2 - 1}}.$$

Thus, we have established:

Proposition 3. *For normal configurations, the disk-director ratio is the eccentricity of the conic, and for abnormal configurations it is the conjugate eccentricity.*

7.4 FOCAL DISK-DIRECTOR THEOREM AND ITS CONVERSE

Because every conic, including a flipped hyperbola, can be obtained as a section of a twisted cylinder, the foregoing results can be summarized as follows:

Theorem 7.1. *For a conic with eccentricity e, there is an infinite family of focal disk-director pairs, all with disk-director ratio equal to e. For a hyperbola there is a second infinite family of focal disk-director pairs, all with disk-director ratio equal to its conjugate eccentricity ε.*

The following converse to Theorem 7.1 shows that conics are the only curves having the focal disk-director property.

Theorem 7.2. (a) *Given a disk, a coplanar line L, and a positive number q, the locus of all points P in the plane of the disk such that the length of the tangent segment from P to the disk is q times the distance from P to L is a noncircular conic. The disk is a focal disk with director L, and q is the disk-director ratio.*

(b) *The conic in* (a) *has eccentricity $e = q$, except when L intersects the focal disk and $q > r/\sqrt{r^2 - \lambda^2}$, where r is the radius of the disk, and $|\lambda| < r$ is the distance from the disk's center to L, in which case $e = q/\sqrt{q^2 - 1}$.*

Proof of (a). In Figure 7.10, choose the origin of xy coordinates at the center of the disk of radius r, and take L parallel to the y axis with equation $x = \lambda$, where $|\lambda| \geq 0$ is the distance from the disk's center to L. The length t of the tangent segment to the disk from point (x, y) on the locus satisfies the defining property

$$t = q\,|x - \lambda|, \tag{7.2}$$

which, together with $x^2 + y^2 = t^2 + r^2$, gives $x^2 + y^2 - r^2 = q^2(x^2 - 2\lambda x + \lambda^2)$ or

$$(1 - q^2)x^2 + y^2 + 2\lambda q^2 x = r^2 + q^2\lambda^2. \tag{7.3}$$

This represents a conic. The disk and L constitute a focal disk-director pair with disk-director ratio q. When $q = 0$, (7.3) represents a circle of radius r.

Proof of (b). $\boldsymbol{q = 1}$**.** In this case, (7.3) represents a parabola (or degenerate case) given by

$$y^2 = -2\lambda x + r^2 + \lambda^2. \tag{7.4}$$

When $\lambda = 0$, the parabola degenerates to the pair of horizontal lines $y^2 = r^2$, tangent to the focal disk as in Figure 7.4c, with the director passing through the center of the disk. If $\lambda \neq 0$, (7.4) gives a nondegenerate parabola with $e = q = 1$.

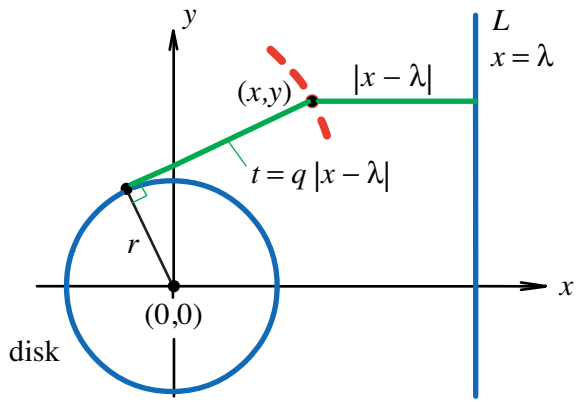


Figure 7.10: Diagram for the proof of Theorem 7.2(a).

$q \neq 1$. By completing the squares in (7.3) we find the equation

$$(x-\rho)^2 + \frac{y^2}{1-q^2} = \frac{B}{1-q^2}, \tag{7.5}$$

which represents a *central conic*, with

$$\rho = -\frac{\lambda q^2}{1-q^2}, \text{ and } B = r^2 - \lambda\rho. \tag{7.6}$$

When $B = 0$, (7.5) becomes $y^2 = (q^2-1)(x-\rho)^2$, which implies $q > 1$ and represents a degenerate case of a hyperbola consisting of two lines intersecting at $(\rho, 0)$ with slopes $\pm\sqrt{q^2-1}$.

Now assume $B \neq 0$, so the central conic is a nondegenerate ellipse or hyperbola. From (7.5) we see that its center is at $(\rho, 0)$ which, by (7.6), is independent of r. To determine the eccentricity, first we find lengths a and b of the semiaxes as follows.

$0 < q < 1$. In this case $B > 0$ and (7.5) can be written as $(x-\rho)^2/a^2 + y^2/b^2 = 1$, which represents a noncircular ellipse with

$$a^2 = \frac{B}{1-q^2}, \text{ and } b^2 = (1-q^2)a^2. \tag{7.7}$$

This gives $q = \sqrt{1 - b^2/a^2}$, which is also the eccentricity, hence $e = q$.

$q > 1$. In this case (7.5) represents a hyperbola given by

$$(x-\rho)^2 - \frac{y^2}{q^2-1} = \frac{-B}{q^2-1}. \tag{7.8}$$

There are two types of hyperbolas, depending on whether $B < 0$ (horizontal focal axis) or $B > 0$ (vertical focal axis).

***B* < 0.** In this case $-B = b^2$ for some $b > 0$ and (7.8) can be written in the form $(x-\rho)^2/a^2 - y^2/b^2 = 1$, where $a^2 = b^2/(q^2-1)$. This hyperbola has its foci on the x axis and eccentricity $e = \sqrt{1 + b^2/a^2} = q$.

***B* > 0.** In this case we write $B = b^2$ for some $b > 0$ and (7.8) takes the form $y^2/b^2 - (x-\rho)^2/a^2 = 1$, where $a^2 = b^2/(q^2-1)$. The foci of this hyperbola are not on the x axis but on the vertical line $x = \rho$. The director $x = \lambda$ is also vertical. This is the flipped hyperbola in the abnormal case. The eccentricity of the flipped hyperbola is $e = \sqrt{1 + a^2/b^2} = q/\sqrt{q^2-1}$. This cannot occur when $r = 0$, which is the classical case, so $B > 0$ requires $r > 0$.

Examples for various q are shown in Figure 7.11. It is easy to verify that the case with $B > 0$ occurs when $\lambda^2 < r^2$ and $q > r/\sqrt{r^2-\lambda^2} = 1/\cos\gamma_0$, where γ_0 is the asymptote angle in Figure 7.11b. Earlier we mentioned that the degenerate

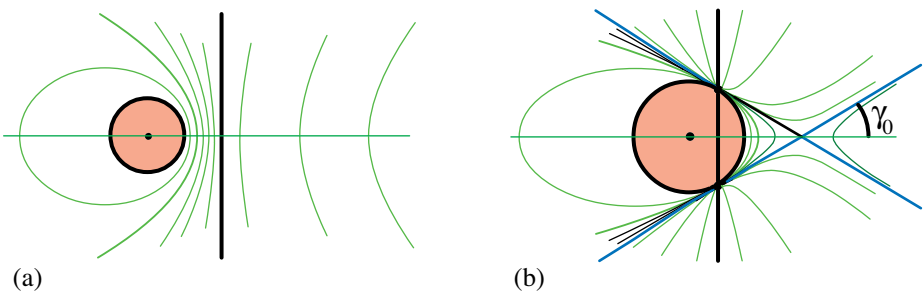


Figure 7.11: Examples of the conic for various values of q. The director intersects the disk in (b) but not in (a).

configuration in Figure 7.4c can be regarded as a limiting case of the abnormal configuration in Figure 7.5c. In terms of eccentricity, this may seem paradoxical because Figure 7.4c is a limit of parabolic configurations, with all parabolas having eccentricity $e = 1$, but as the configuration in Figure 7.5c approaches that in Figure 4c, the eccentricities of the hyperbolas tend to ∞. However, there is no paradox in terms of the disk-director ratio q because during this limit process $q \to 1$, which matches the value of q in Figure 7.4c.

PART 2: BIFOCAL DISK DESCRIPTION OF CONICS

7.5 BIFOCAL DISK PROPERTY

Return to the twisted cylinder cut by an inclined plane (Figure 7.2), and inscribe in it two spheres as in Figure 7.12, each of which pierces a focal disk in the plane of the conic. We will prove the following *bifocal disk property.*

Proposition 4. *There is a constant c such that for each point P on the conic of intersection, either the sum or absolute difference of the tangent lengths from P to the two focal disks is equal to c.*

Proof. In Figure 7.12, ellipses are shown, but the argument applies to all conics, including circular cross sections.

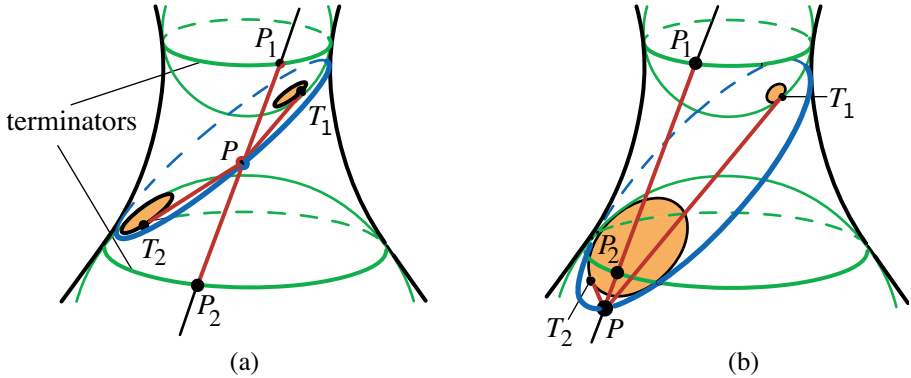


Figure 7.12: Bifocal disk property:(a) $PT_1 + PT_2 = c$. (b) $PT_1 - PT_2 = c$.

A generator of the twisted cylinder passing through a point P on the conic intersects the terminator of sphere 1 at P_1 and that of sphere 2 at P_2. The tangent segment PT_1 from P to disk 1 has the same length as the tangent segment PP_1 because both are tangent to sphere 1. Similarly, $PP_2 = PT_2$ for sphere 2. Therefore, if P is between the upper and lower terminators as in Figure 7.12a, we have the sum

$$PT_2 + PT_1 = PP_2 + PP_1 = P_2P_1 = c,$$

where $c = P_2P_1$ is the constant distance between the terminators measured along the generator on the twisted cylinder. But if P lies below the lower terminator so that $PT_2 < PT_1$ as in Figure 7.12b, the same argument shows that the difference

$$PT_1 - PT_2 = PP_1 - PP_2 = P_2P_1 = c,$$

where c is the same constant distance above. Therefore, the sum or absolute difference of the tangent lengths (larger minus smaller) from a point on the conic of intersection to the two focal disks is constant.

Possible configurations.

Figures 7.13 and 7.14, elaborations of Figure 7.3, show examples of some of the configurations of conics and two focal disks (shown shaded) that can occur. The directors are also shown as intersections of the cutting plane and the two terminator planes.

In Figure 7.13c the cutting plane has been translated to produce a degenerate parabola (two parallel lines). In Section 7.13 we show that this degenerate parabola leads to a surprising property of hyperbolas! Figure 7.14 relates focal disks to the hyperbolic cross sections shown in Figure 7.3c. An abnormal case is depicted in Figure 7.14c. Unlike the normal cases in Figures 7.13a, 7.13b, and 7.14a, the disks in Figure 7.14c cannot be shrunk to become foci.

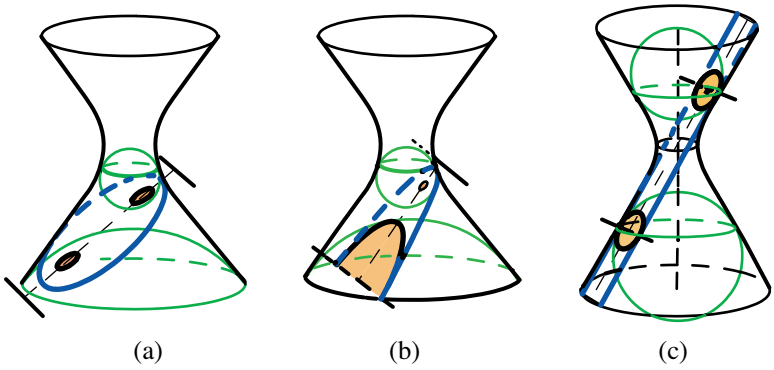


Figure 7.13: (a) Ellipse. (b) Parabola. (c) Degenerate parabola obtained by translating the cutting plane in (b). This degenerate conic cannot be obtained as a section of a cone.

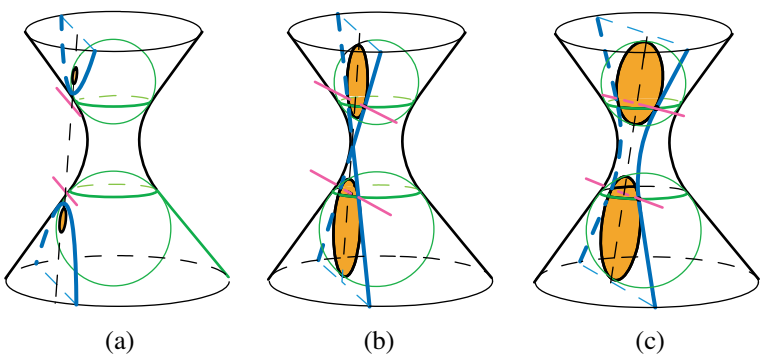


Figure 7.14: (a) Hyperbola. Translating the cutting plane in (a) produces a degenerate hyperbola (b). Further translation leads to the abnormal configuration in (c).

By moving the inscribed spheres independently, we obtain:

Proposition 5. *Every conic of intersection has the bifocal disk property with respect to infinitely many pairs of focal disks.*

Geometric relations on a twisted cylinder.

Figure 7.15 relates geometric parameters of the inscribed spheres, terminators, focal disks, and directors with the angles α and β introduced in Figure 7.7. Figure 7.15a shows that

$$c = s\cos\alpha, \tag{7.9}$$

where s is the distance between the centers of the inscribed spheres, and c is the length of the portion of the generator joining the terminators. Figure 7.15b gives

$$h = c\cos\alpha, \tag{7.10}$$

where h is the distance between the terminator planes. In Figure 7.15c,

$$d = s \sin \beta, \tag{7.11}$$

where d is the distance between centers of the focal disks. Note that $d = 0$ when $\beta = 0$, which means the focal disks are concentric when the cutting plane is horizontal. For $\beta > 0$ Figure 7.15d yields

$$h = D \sin \beta, \tag{7.12}$$

where D is the distance between the parallel directors of the focal disks.

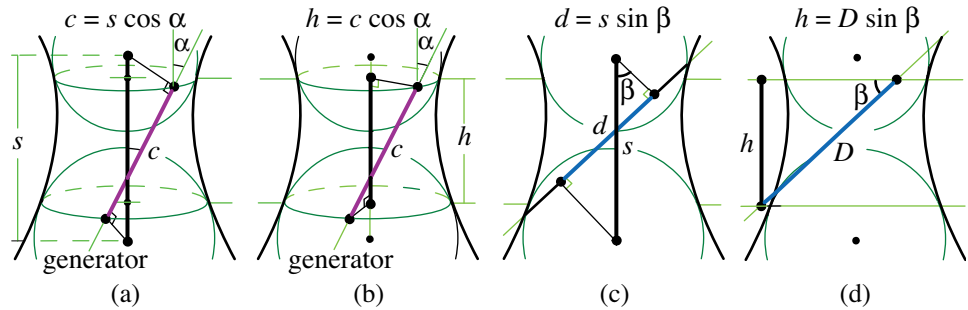


Figure 7.15: Relating parameters c, s, h, d, D with angles α and β.

Dividing (7.11) by (7.9) and using (7.1) we find

$$\frac{d}{c} = \frac{\sin \beta}{\cos \alpha} = q. \tag{7.13}$$

Because q is related to the eccentricity of a noncircular conic as described in Proposition 3, (7.13) shows that the ratio d/c bears the same relation to eccentricity in the focal disk-director property.

Proposition 6. *If $d > 0$, the ratio d/c of the distance d between centers of the focal disks and the constant sum or absolute difference c of tangent lengths is equal to the disk-director ratio q of the configuration.*

Equating (7.10) and (7.12) and using(7.13) we also deduce

$$D = \frac{d}{q^2}, \tag{7.14}$$

which relates the distance D between directors to the distance d between centers of the focal disks.

Possible pairs of focal disks.

Consider a fixed ellipse and its family of focal disks in Figure 7.4a. Any two disks in this family can serve as a pair of focal disks. If the ellipse has eccentricty e,

major axis of length $2a$, and distance $2f$ between foci, there are infinitely many pairs of positive constants c and d with ratio $d/c = e$ (because $q = e$) subject to the constraints $c \leq 2a$ and $d \leq 2f$, the same restrictions imposed by the inscribed spheres in Figure 7.12. If one permissible value of c or d is chosen, the other is determined by the relation $d/c = e$, and each such pair corresponds to a pair of focal disks. In Figure 7.4a, the solid line joining the foci is the locus of centers of the focal disks.

For a parabola ($q = e = 1$) there are infinitely many choices of positive constants c and d with $c = d$ (with no constraints), hence infinitely many pairs of focal disks (overlapping as well as nonoverlapping) with centers at a distance d apart and with a constant sum or absolute difference c of tangent lengths. Examples are shown in Figure 7.4b.

Figure 7.5 shows that there is more than one way for pairs of focal disks to relate to a hyperbola. For normal configurations as in Figure 7.5a, with $q = e = d/c$, one disk can be chosen inside each branch, or both disks can be chosen inside the same branch to form a pair of focal disks. Disks inside the same branch have no further constraints on c or d, but disks inside different branches require $d > 2f$, $c > 2a$. For abnormal configurations as in Figure 7.5c, any two disks tangent to both branches of the hyperbola can be chosen as a pair of focal disks with no constraints on c or d except $d/c = q$, where now $q = \varepsilon$, the conjugate eccentricity of the hyperbola.

Tandem motion.

Relations (7.9)-(7.14) reveal interesting phenomena not otherwise immediately apparent. The inscribed spheres in Figure 7.15 can be moved in tandem, that is, with fixed distance s between their centers. Their radii will change. The pierced disks in the cutting plane will also move and, by (7.11), the distance d between their centers remains fixed, so the disks also move in tandem in the plane of the conic, and their radii will also change. The radii of the terminators will also change but, because α is constant, the length c of the generator joining the terminators, and the distance h between planes of the terminators do not change (by (7.9) and (7.10)). Because c does not change, the bifocal disk property holds with the same value of c during the entire tandem motion. By (7.12), the distance D between directors does not change.

7.6 BIFOCAL DISK THEOREM AND ITS CONVERSE

Every conic, including a flipped hyperbola, can be obtained as a section of a twisted cylinder. In view of the foregoing remarks we have the following bifocal disk theorem:

Theorem 7.3. *For every conic and each permissible value c of constant sum or absolute difference of tangent lengths, there is an infinite family of pairs of focal disks satisfying the bifocal disk property for c.*

We remind the reader that for abnormal hyperbolic configurations, the family of focal disks in Figure 7.5c is not realizable on a cone.

The following converse of Theorem 7.3 extends the classical bifocal property of central conics [1, p. 498] to all conics, including the parabola. It also tells us that conics are the only curves having the bifocal disk property. Therefore this property characterizes the conics.

Theorem 7.4. *Suppose we are given two coplanar disks with distance $d > 0$ between their centers. The locus of all points in the plane such that either the sum or the absolute difference of the lengths of the tangent segments from this point to the two disks is a positive constant c is a conic. Each disk is also a focal disk with a director having disk-director ratio $q = d/c$, regardless of the radii of the disks. The eccentricity of the conic is related to q as described in* Theorem 7.2b.

Before presenting the proof, we note that the locus problem described in Theorem 7.4 involves four parameters (the two radii of the disks, the distance d between their centers, and the constant c). Nevertheless, our proof will reduce this to the locus problem of Theorem 7.2, which involves only three parameters: the radius of one disk, the distance to its director, and the disk-director ratio.

Proof. Place the center of one disk of radius r_1 at the origin and the center of the other disk of radius r_2 at $(d, 0)$, where $d > 0$.

Denote the lengths of the tangent segments from a point (x, y) on the locus by t_1 and t_2, as in Figure 7.16. Then (x, y) is on the locus if and only if either $t_1 + t_2$

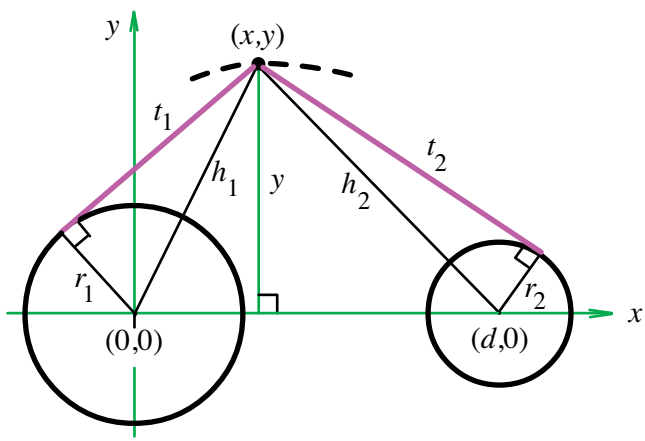


Figure 7.16: Diagram for the proof of Theorem 7.4.

or $t_1 - t_2$ is either c or $-c$. This holds if and only if either

$$(c - (t_1 + t_2))(c - (t_1 - t_2)) = 0$$

or

$$(c + (t_1 + t_2))(c + (t_1 - t_2)) = 0.$$

This is equivalent to

$$c^2 \pm 2ct_1 + (t_1^2 - t_2^2) = 0. \tag{7.15}$$

Right triangles with legs t_1 and t_2 reveal that $t_1^2 = h_1^2 - r_1^2$ and $t_2^2 = h_2^2 - r_2^2$, so

$$t_1^2 - t_2^2 = h_1^2 - h_2^2 + r_2^2 - r_1^2. \tag{7.16}$$

But h_1 and h_2 are also hypotenuses of right triangles with altitude y, so

$$h_1^2 = x^2 + y^2, \quad h_2^2 = (x-d)^2 + y^2,$$

giving $h_1^2 - h_2^2 = 2dx - d^2$, which transforms (7.16) into

$$t_1^2 - t_2^2 = 2dx - d^2 + r_2^2 - r_1^2.$$

Use this in (7.15) and solve for t_1 to obtain

$$t_1 = \pm(\frac{d}{c}x + \frac{k}{2c}),$$

where

$$k = c^2 - d^2 + r_2^2 - r_1^2. \tag{7.17}$$

But $t_1 \geq 0$, so

$$t_1 = \frac{d}{c}|x - \lambda_1|, \tag{7.18}$$

where

$$\lambda_1 = -\frac{k}{2d}. \tag{7.19}$$

At this stage the locus problem has been reduced to that in Theorem 7.2a. Comparing (7.18) with the defining property (7.2) and applying Theorem 7.2a we see that the locus is a conic. The disk of radius r_1 is also a focal disk having the line $x = \lambda_1$ as its inherited director and with disk-director ratio $q = d/c$. The cartesian equation of the conic is given by (7.3) with $\lambda = \lambda_1$ and $r = r_1$. The eccentricity of the conic is described in terms of q by Theorem 7.2b.

By an argument similar to that leading to (7.18), it is easy to verify that the disk of radius r_2 is also a focal disk for exactly the same conic, with its own inherited director $x = \lambda_2$ and the same disk-director ratio $q = d/c$, where

$$\lambda_2 = \lambda_1 + \frac{d}{q^2}. \tag{7.20}$$

Moreover, (7.20) shows that the directors are parallel and the distance between them is d/q^2, in agreement with (7.14). The directors are always orthogonal to the line through the centers of the disks. Examples illustrating this theorem are given in Section 7.10.

Theorem 7.4 assumes $d > 0$ and does not cover the case of concentric focal disks. But in this case it is easy to verify directly that the locus in question is a concentric circle containing the disks. In fact, each of the tangent distances t_1 and t_2 from that circle to the focal disks is constant, so the sum of tangent distances is equal to the constant $t_1 + t_2$ and the absolute difference is the constant $|t_1 - t_2|$. Also,

Theorem 7.4 does not cover the case $c = 0$ (equal tangent lengths), which is easily analyzed directly and is illustrated in the last diagram in Figure 7.22.

The solution of the four-parameter problem in Theorem 7.4 leads to the same set of conics as does the solution of the three-parameter problem in Theorem 7.2. This means that as we vary the four parameters in Theorem 7.4 we obtain conics from the same set with repetitions. This is consistent with the converse result in Theorem 7.3, which provides an infinite family of pairs of focal disks for a given conic.

7.7 TWO NEW CHARACTERIZATIONS OF THE CONICS

The results obtained in Theorems 7.1 through 7.4 can be summarized into one theorem that provides the two new characterizations of the conics mentioned in Section 7.1.

Theorem 7.5. *The focal disk-director property is satisfied by all conics, and only by the conics. The bifocal disk property is satisfied by all the conics, and only by the conics.*

A central conic has two foci and two directrices, but a parabola has one focus and one directrix. By contrast, we have shown that every conic, including the parabola, has infinitely many pairs of focal disks, and every conic has infinitely many focal disk-director pairs.

PART 3: SUPPLEMENTARY RESULTS

7.8 MORE ON DIRECTORS AND THE BIFOCAL DISK PROPERTY

If a director intersects the boundary of the focal disk at P, then in Figures 7.8a and 7.8c, $PT = q\,PD = 0$ trivially, so P is on the conic. These intersections separate the conic into complementary portions, on one of which the sum of tangent lengths is constant, while on the other the absolute difference of tangent lengths is the same constant. Next we show how the directors identify the complementary portions.

Proposition 7. *Suppose we are given a conic having two focal disks with distance $d > 0$ between their centers. Let q be the disk-director ratio for each disk-director pair. Then on the portion of the conic between directors the sum of tangent lengths to the disks has the constant value $c = d/q$, and on the remaining portions the absolute difference of tangent lengths is the same constant.*

Proof. If P is on the conic with tangents to the focal disks of lengths t_1 and t_2, the distances from P to the corresponding directors are t_1/q and t_2/q. By (7.14), the distance between directors is d/q^2. Therefore, if P lies between the directors, then $t_1/q + t_2/q = d/q^2$, so $t_1 + t_2 = d/q = c$. Otherwise, $|t_1/q - t_2/q| = d/q^2$, in which case $|t_1 - t_2| = d/q = c$.

Special cases of Proposition 7.

The classical bifocal properties of central conics are special cases of Proposition 7 for normal configurations. To see why, keep d fixed and let the focal disks shrink to the foci F_1 and F_2, so the directors become directrices. Every point P on an ellipse lies between its two directrices, so by Proposition 7 the sum $PF_1 + PF_2$ is constant. Points on a hyperbola cannot lie between the two directrices, and Proposition 7 tells us that the absolute difference $|PF_1 - PF_2|$ is constant on the entire hyperbola.

What happens if both disks are very near to one focus? Figure 7.17 shows examples with one focal disk inside another. In Figure 7.17b, one of the disks is itself a focus. In these examples no portion of the conic is between the directors, revealing the surprising fact that the difference of tangent lengths can be constant everywhere on an ellipse and also on a parabola. This resembles the classical bifocal property of a hyperbola.

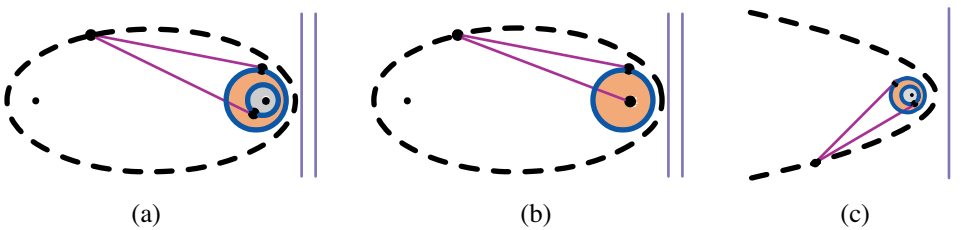


Figure 7.17: The difference of tangent lengths is constant everywhere on the ellipse in (a) and (b), and also everywhere on the parabola in (c).

7.9 LOCATING A FOCAL DISK AND ITS DIRECTOR FOR A CONIC

Next we address the following problem related to the families of focal disks in Theorems 7.1 and 7.3:

For a noncircular conic, determine the center and radius of a possible focal disk.

We solve this problem using formulas in the proof of Theorem 7.2b to obtain, for each type of conic, a relation between the radius r and center of a focal disk. The central conics (eccentricity $e \neq 1$) are treated separately from the parabola ($e = 1$). In both cases we take the x axis as an axis of symmetry passing through the foci.

Central conic.

According to the proof of Theorem 7.3b, the center of a nondegenerate central conic is at $(\rho, 0)$ where ρ is related to λ by (7.6), which gives $\lambda = \rho - \rho/q^2$, and the length a of the semimajor axis is given by (7.7), which implies

$$a^2 = \frac{r^2}{1 - q^2} + \frac{\rho^2}{q^2}.$$

The parameter ρ is positive or negative, and $|\rho|$ represents the distance between the center of the focal disk and the center of the conic. Therefore, if the central conic is translated horizontally so its center is at the origin and the focal disk is at the point $(-\rho, 0)$, then the director is translated to the line $x = \lambda - \rho = -\rho/q^2$, with the focal disk and director on the same side of the origin. The formula for a^2 remains unchanged because ρ^2 and r^2 are unchanged.

For a given a and disk-director ratio $q \neq 1$, introduce $f = aq$, the distance from the center of the conic to a focus, and rewrite the foregoing formula for a^2 in the form

$$\frac{\rho^2}{f^2} + \frac{r^2}{a^2(1-q^2)} = 1. \tag{7.21}$$

Now we take a central conic with center at the origin and semiaxes a and b, where $b^2 = a^2|1-q^2|$, and determine a possible focal disk-director pair for it. Because the directrix is the line $x = -\rho/q^2$, we wish to determine all pairs ρ and r satisfying (7.21). For an ellipse, $q < 1$, $b^2 = a^2(1-q^2)$ and (7.21) becomes

$$\frac{\rho^2}{f^2} + \frac{r^2}{b^2} = 1, \tag{7.22}$$

which implies $|\rho| \leq f$ and $r \leq b$.

For a hyperbola of Type 1, ($B < 0$ in (7.8)) we have $q > 1, b^2 = a^2(q^2-1)$, and (7.21) takes the form

$$\frac{\rho^2}{f^2} - \frac{r^2}{b^2} = 1, \tag{7.23}$$

which implies $|\rho| \geq f$.

A hyperbola of Type 2 (given by (7.8) with $B > 0$), when translated horizontally by $-\rho$, has its vertices at $(0, \pm b)$ on a vertical axis through the center of the hyperbola where $b > 0$ and b^2 is given by (7.17). The foci of the hyperbola are also on the vertical line, but its focal disks have their centers on the x axis. In this case (7.21) relates the parameters ρ and r by the equation

$$\frac{r^2}{b^2} - \frac{\rho^2}{f^2} = 1, \tag{7.24}$$

where now $r \geq b$.

Parabola.

In this case the relation between ρ and r is obtained differently. For the parabola in (7.4) the vertex is at $(\rho, 0)$, $\lambda = -2f$, and $\rho + f = r^2/(2\lambda) = -r^2/(4f)$, so $r^2 = -4f(\rho + f)$. This can be written as

$$r^2 = 4f(-\rho - f), \tag{7.25}$$

and it implies $|\rho| \geq f$. In the ρr plane, (7.25) represents a parabola with its vertex at $(f, 0)$, where f is the distance from vertex to focus. In this case, the corresponding director is $x = \rho - 2f$.

The results of this section can be summarized as follows.

Theorem 7.6. (a) *For a central conic with center at the origin and with disk-director ratio q, a focal disk with center at $(\rho, 0)$ has radius given by* (7.22)-(7.24), *and its director is along the line $x = \delta$, where*

$$\delta = \frac{\rho}{q^2}. \tag{7.26}$$

(b) *For a parabola that opens to the right, with vertex at the origin and focus at $(f, 0)$, a focal disk with center at $(\rho, 0)$ has radius given by* (7.25), *and its director is the line $x = \delta$, where*

$$\delta = \rho - 2f. \tag{7.27}$$

Geometric construction of focal disk.

In the ρr plane, (7.21) represents a central conic of the same type as the given conic with its center at the origin. This auxiliary conic is a horizontal dilation by the factor q of the conic with semiaxes a and b. Now we show that the auxiliary conic can be used to construct every focal disk of the central conic geometrically.

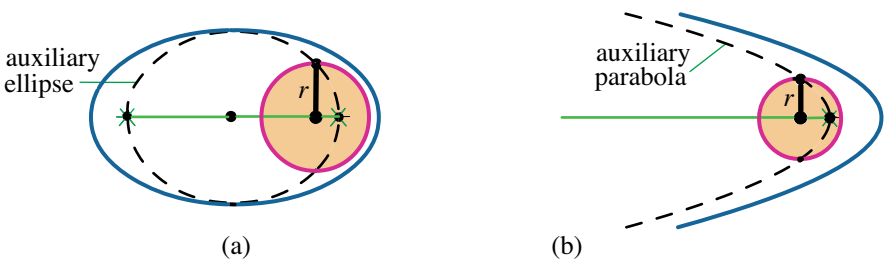


Figure 7.18: Auxiliary (dashed) conic used to construct a focal disk, for (a) an ellipse, and (b) a parabola.

The construction is illustrated in Figure 7.18a for an ellipse, where the auxiliary ellipse given by (7.22) (shown dashed) has its vertices at the foci of an ellipse with semiaxes a and b. The auxiliary ellipse is a horizontal dilation by the factor q. A point (ρ, r) on the auxiliary ellipse satisfying (7.22) lies directly above the center of the focal disk for the given ellipse, and its ordinate r is the radius of the disk. We obtain all focal disks for the ellipse by allowing ρ to vary from $-f$ to f. The same method produces the focal disks of a hyperbola with semiaxes a and b. Horizontal dilation by the factor e produces the hyperbolas in (7.23) and (7.24), which can be used in a similar way to construct the focal disks of the hyperbola. In Figure 7.18b the dashed auxiliary parabola described by (7.25) is obtained by translating the parabola in (7.4) so the translated vertex matches the focus of the given parabola.

How do we locate the director for a given focal disk?

This is answered by (7.20) which can be interpreted as the following:

Shifting principle. *If one focal disk of a family is shifted through a distance d to another, its director will be shifted in the same direction by the distance d/q^2.*

In the abnormal case in Figure 7.5c we can start with the smallest disk whose director passes through its center, and shift it by the distance d to another disk of the family. The director is then shifted in the same direction by d/q^2. For the families in Figures 7.4a, 7.4b, and 7.5a, we can start with a focus as one focal disk, and shift it by d to another disk of the family. The directrix for that focus is then shifted by the distance d/q^2 to locate the director of the other disk.

7.10 EXAMPLES OF CONICS WITH FIXED FOCAL DISKS

This section fixes the focal disks and provides a series of snapshots showing what happens qualitatively to the conics as c varies. In Figures 7.19-7.22 we take $r_1 = 2, r_2 = 1, d = 6$ and let c decrease from $c = 10$ to $c = 0$. The arrows indicate the positions of the directors. Between directors the *sum* of tangent lengths to the disks is constant, as indicated by the solid curves. On the remaining portion of the conic

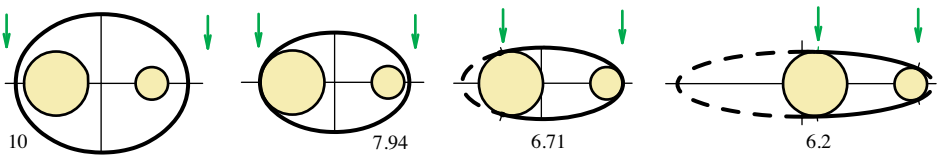


Figure 7.19: Ellipses become more elongated as c decreases from 10 toward 6.

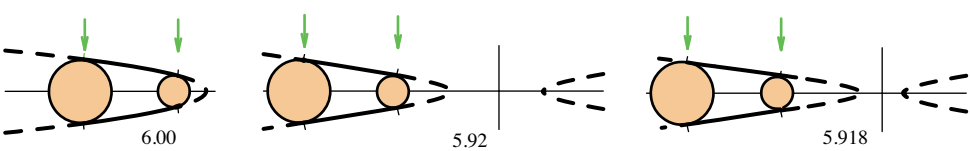


Figure 7.20: Conic becomes a parabola when $c = 6$, and a hyperbola when $c < 6$.

the *absolute difference* of tangent lengths is constant, as indicated by the dashed curves. If c is large, the ratio d/c is close to zero, and the conic resembles a circle of radius $c/2$. As c decreases, the ratio d/c increases and the ellipse becomes more and more elongated (Figure 7.19). When $c = d$, the ellipse suddenly changes to a parabola, and then to a hyperbola as c continues to decrease to 0, as shown in Figures 7.20 through 7.22.

When one disk is a point (a focus), the two extreme degenerate cases in Figure 7.21 coincide, and the abnormal configurations between them disappear. In this case, all possible configurations can be realized as sections of a cone, and the twisted cylinder is not needed. It is important to note that in this sequence of snapshots the disk-director ratio q increases monotonically, but the eccentricity e is not monotonic. The eccentricity jumps up when the hyperbola flips in Figure 7.21, and then jumps down when the hyperbola flips back in Figure 7.22. This is consistent with the fact that $e = q$ in normal cases, and $e = q/\sqrt{q^2 - 1}$ in abnormal cases.

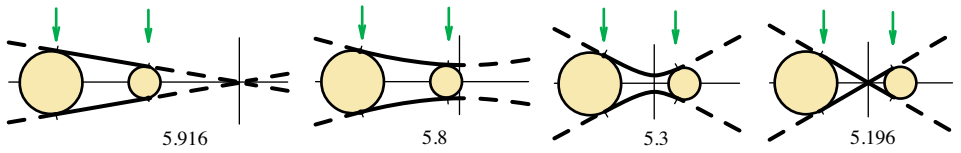


Figure 7.21: The hyperbola degenerates to a pair of lines when $c = \sqrt{d^2 - (r_1 - r_2)^2}$. As c decreases further the hyperbola flips and opens up and down instead of right and left, until the next degeneration occurs at $c = \sqrt{d^2 - (r_1 + r_2)^2}$. This interval for c represents abnormal configurations not realizable on a cone, but requiring a twisted cylinder.

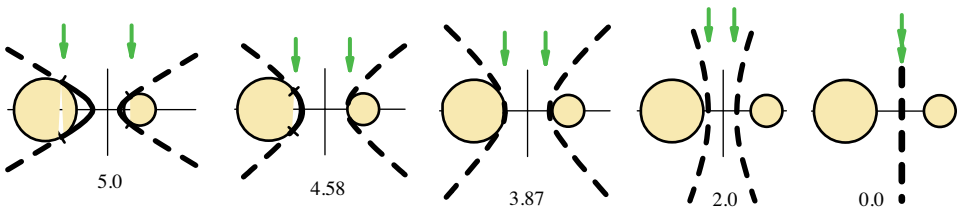


Figure 7.22: For $c < \sqrt{d^2 - (r_1 + r_2)^2}$ the hyperbola flips back with each branch tangent to one of the disks, until $c < \sqrt{(d - r_1)^2 - r_2^2}$, when both disks detach. As c approaches 0, both branches tend to one vertical line, on which the lengths of the tangents to the two disks are equal: $t_1 = t_2$.

7.11 APPLICATIONS OF THE BIFOCAL DISK PROPERTY TO TRACING CONICS

Known string mechanisms for tracing conics, based on their classical focal properties, require separate devices for the ellipse, hyperbola, and parabola, as shown in Figure 7.23. The bifocal disk property allows us to design two single mechanisms, each of which traces all three types.

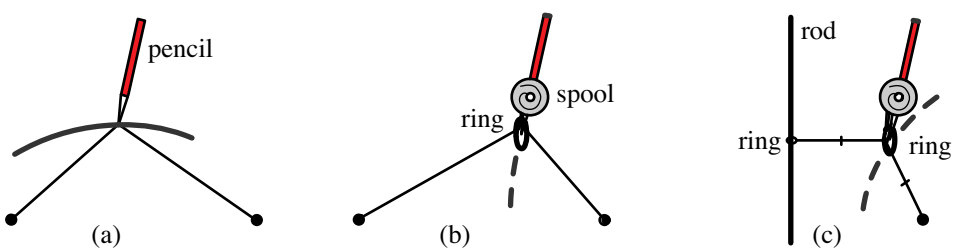


Figure 7.23: Known mechanisms for tracing (a) ellipse, (b) hyperbola, (c) parabola.

Figure 7.24a (insert) shows a plus-shaped receptacle that accommodates two perpendicular rigid rods, one for the radius of the disk, the other for the tangent line to the disk. A thumb screw allows the radius to be changed. Two such receptacles are needed, one for each focal disk. The two tangent rods pass through a small ring

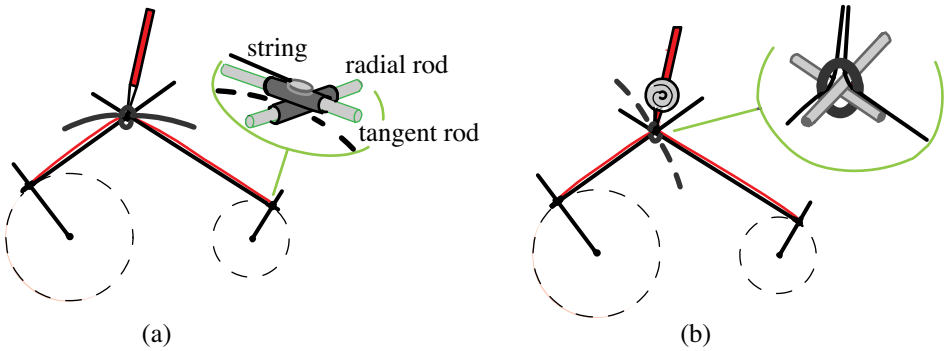


Figure 7.24: Two new mechanisms, each based on the bifocal disk property, for tracing portions of all three conics. In (a) the sum of tangent lengths is constant, while in (b) the difference of tangent lengths is constant.

at their intersection. A string passing through the ring connects one receptacle to the other, as in Figure 7.24a. The string has constant length, so a pencil through the ring will trace out that part of the conic on which the sum of the tangent lengths is the constant length. By changing only the length of the string we can achieve any eccentricity to obtain portions of all three types of conics. Thus, with the single mechanism in Figure 7.24a we can trace all the solid portions of the conics shown in Section 7.5. Larger solid portions can be traced by changing the radii of the focal disks and the distance between their centers.

Figure 7.24b shows another single mechanism that exploits the bifocal disk property that the difference of tangent lengths can be constant everywhere on the conic. Two strings attached to the receptacles pass through the ring and together wind around a spool attached to the pencil. This traces out the dashed portions of the conics shown in Section 7.5, and, what is more impressive, it also traces the complete conic of any eccentricity. In this case we can achieve any eccentricity by changing the length of one string between the disk and the ring. Examples for an ellipse and parabola are shown in Figure 7.23.

7.12 FOCAL DISKS AND DIRECTORS FOR THE ELLIPSE AS A SECTION OF A CIRCULAR CYLINDER

Every ellipse is a cross section of a right circular cylinder cut by an inclined plane, as illustrated in Figure 7.25a. Because a circular cylinder is a special case of a twisted cylinder, focal disks and directors for an ellipse also occur naturally in such cross sections. The analysis on a circular cylinder is somewhat simpler and is presented here primarily because this special case led to our discovery of the bifocal disk property for all conics.

Two spheres inscribed in the cylinder touch the cylinder along parallel equators. In Figure 7.25a, the spheres pierce the cutting plane along two circular disks. In Figure 7.25b, each disk has radius 0 and the spheres are tangent to the cutting plane at F_1 and F_2, the foci of the ellipse. In this case the diagram in Figure 7.25c reduces

to the classical bifocal property for an ellipse: *the sum $PF_1 + PF_2$ is constant.*

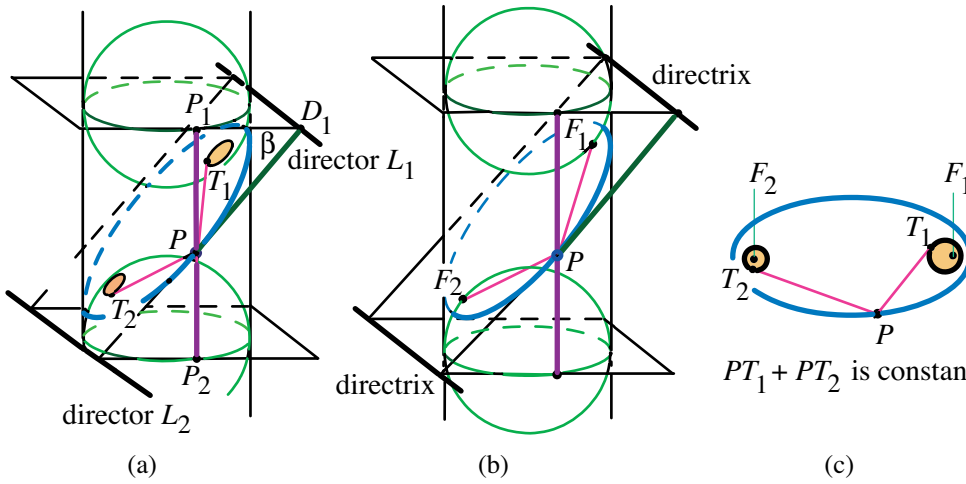


Figure 7.25: (a) Cutting plane cuts focal disks from inscribed spheres. (b) Cutting plane tangent to the inscribed spheres. (c) The sum of tangent lengths to focal disks is constant.

First we show that the disks satisfy the bifocal disk property in Figure 7.25c, and then we show that they are focal disks whose directors are the parallel lines of intersection L_1 and L_2 of the equatorial planes with the cutting plane.

To verify the bifocal disk property, refer to Figure 7.25a, and let P be a point on the ellipse. A vertical line on the cylinder through P intersects the equators of the inscribed spheres at P_1 and P_2, with PP_1 tangent to the upper sphere and PP_2 tangent to the lower sphere. Let PT_1 and PT_2 denote the lengths of the tangent segments from P to the upper and lower disks, respectively.

The two tangent segments from P to a disk and to the corresponding equator have equal lengths, hence $PT_1 = PP_1$ and $PT_2 = PP_2$. But the two vertical tangents from the ellipse to the equators have lengths of sum $PP_1 + PP_2 = P_1P_2$, which is the distance between the parallel equators. This sum is a constant (independent of P) because of the rotational symmetry of the cylinder, so the sum of the tangent segments to the disks is the same constant:

$$PT_1 + PT_2 = P_1P_2.$$

Moreover, this relation holds if we move the spheres relative to the cutting plane, provided neither equator touches the ellipse. In other words, if neither sphere touches the ellipse, as in Figure 7.25, the sum of the lengths of the tangents from the ellipse to the two focal disks is constant. When a sphere touches the ellipse, a dramatic change occurs, as revealed below. Before analyzing this case, we show that the two disks are, in fact, focal disks with directors L_1 and L_2.

In Figure 7.25a, let PD_1 denote the distance from P to the line of intersection L_1 of the cutting plane and the upper equatorial plane. Then the disk-director ratio

$q = PT_1/PD_1 = PP_1/PD_1 = \sin\beta$, where β is the angle the cutting plane makes with the horizontal plane. Because this angle is constant (independent of P) the ratio PT_1/PD_1 is constant on the entire ellipse. This proves that L_1 is the director associated with focal disk 1 and with disk-director ratio $q = \sin\beta$, which is constant on the ellipse. When disk 1 shrinks to F_1, the director L_1 is the classical directrix associated with F_1, and with eccentricity $e = \sin\beta$.

Similarly, L_2 is the director for disk 2 and it becomes the classical directrix for focus F_2 when disk 2 shrinks to F_2. Because $q = \sin\beta$, this shows that the disk-director ratio for an ellipse is equal to its eccentricity: $q = e$.

Bifocal disk property.

Figure 7.26 shows changes that occur in the plane of the ellipse in Figure 7.25 as the upper sphere moves down continuously while the lower sphere is kept fixed. The upper focal disk starts as the focus F_1 and gradually increases in radius, staying inside the ellipse as in Figure 7.26a, until it reaches a critical position where it first touches the ellipse at a vertex, as shown in Figure 7.26b. At that point, the focal disk is tangent to its director, the equator, and the ellipse. As the sphere

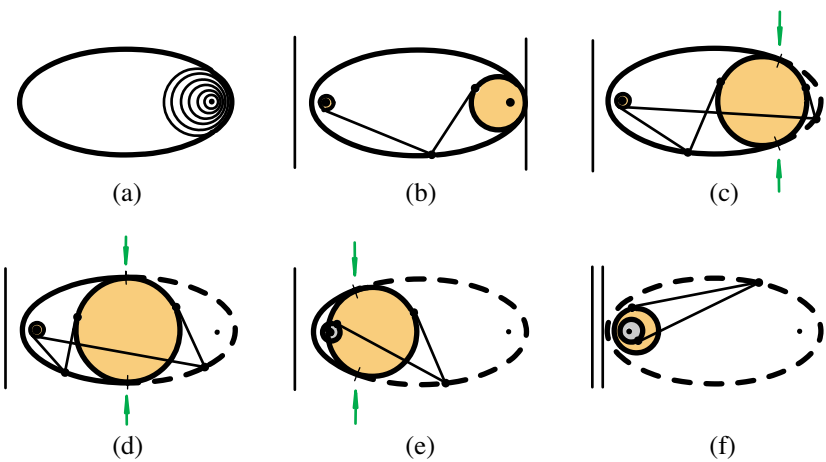


Figure 7.26: Positions of upper focal disk as upper sphere moves down. The focal disk lies inside the ellipse as in (a) until it first touches the ellipse as in (b). In (c)-(e), the focal disk is tangent to the ellipse at two points, and the sum of tangent lengths is constant on the left part of the ellipse, but the difference of tangent lengths is the same constant on the right part, indicated by dashed curves. The arrows indicate how the director separates the two parts. In (f) the disk no longer touches the ellipse, and the difference is constant on the entire ellipse.

moves further down from this critical position, the focal disk touches the ellipse at two points but moves to the left, as in Figures 7.26c, 7.26d, and 7.26e, and a new property is revealed: For those points on the ellipse above the upper sphere's equator, indicated by dashed curves in Figures 7.26c-7.26e, the difference of tangent

lengths is equal to the constant distance P_1P_2 between equators, but for those below that equator, indicated by solid curves, the sum is P_1P_2. In each case the director passes through the points of tangency, as indicated by the arrows.

When the sphere moves down even further and is no longer tangent to the ellipse, the difference of tangent lengths is constant on the entire ellipse, drawn dashed in Figure 7.26f.

Thus, we have the *bifocal disk property*:

On the portion of the ellipse between the two directors, the sum of tangent lengths to the two focal disks is constant, while on the remaining portion of the same ellipse the absolute difference is the same constant.

Of course, a similar situation prevails if the upper sphere is fixed and the lower sphere moves upward. We obtain a symmetrically located set of focal disks. In fact, there is no need to use two spheres, one above and another below the cutting plane. We can construct all possible focal disks by allowing one sphere to move freely upward or downward through the cutting plane.

Profile view of the focal disks.

Using one inscribed sphere instead of two makes it easier to visualize how the focal disks change. We can also keep the inscribed sphere fixed and translate the cutting plane upward (parallel to itself) to get congruent ellipses. The sloping parallel lines in Figure 7.27a show five profiles of the cutting plane as it moves up from the point of tangency (a focus) at the bottom line until it cuts the center of the sphere at the

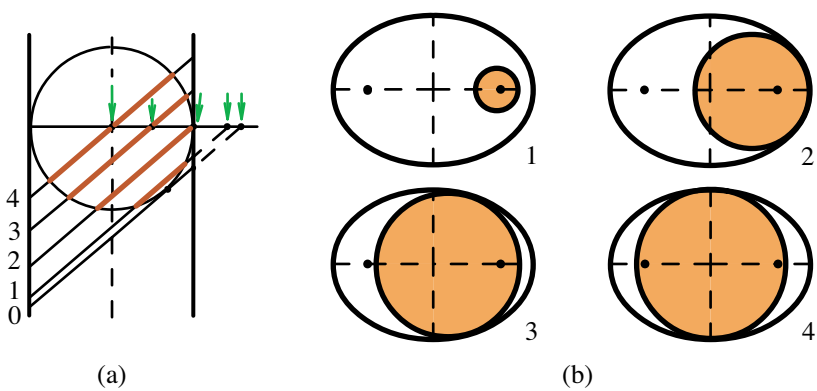


Figure 7.27: (a) Profile view showing relative positions of the cutting plane and the inscribed sphere. Darker segments are profiles of the focal disks in (b).

top line. The darker segments are profiles of the corresponding focal disks, shown shaded in Figure 7.27b. (The focus itself moves along a vertical line.) The line through the sphere's center orthogonal to the cutting plane is the locus (not shown)

of the centers of the focal disks. The arrows show the locations of the directors in profile view.

7.13 SURPRISING PROPERTY OF HYPERBOLAS

The degenerate parabola in Figure 7.13c is shown again in Figure 7.28, which provides a proof without words of the following surprising property of hyperbolas:

For each member of the family of hyperbolas with the same pair of vertices, every circle tangent to both branches intersects each asymptote along a chord of constant length, regardless of the location of the circle and the angle between the asymptotes. This constant length is the distance between vertices.

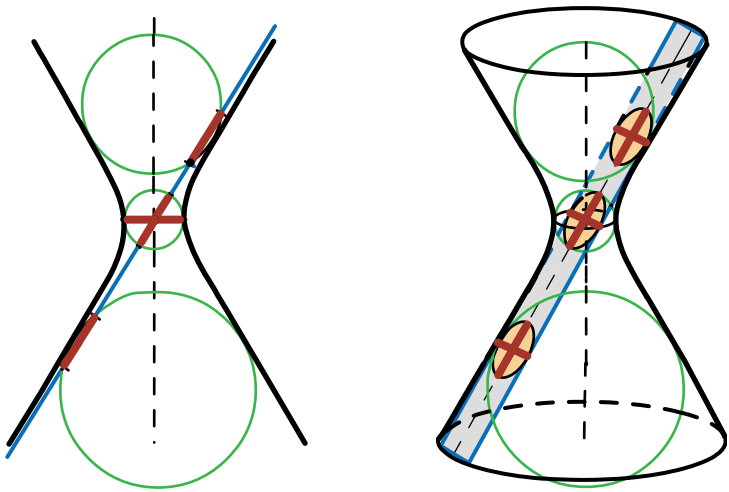


Figure 7.28: The diagram on the left illustrates the property of hyperbolas, and the diagram on the right shows why it is true.

NOTES ON CHAPTER 7

To the best of our knowledge, no one previously used cross sections of twisted cylinders to completely characterize the conic sections. Dandelin [36] used the hyperboloid of revolution to give new proofs of the Theorem of Pascal on hexagons inscribed in conic sections, and of the Theorem of Brianchon on hexagons circumscribing conic sections. As preparation he showed that every conic is a cross section of a hyperboloid cut by a plane, by inscribing spheres whose points of tangency with the cutting plane represent the foci of the conic, as he had done earlier on a cone [1, p. 498]. But in his treatment of the hyperboloid (and of the cone) he did not pierce the inclined plane with the spheres as we have done to produce focal disks and directors.

Some of our results were anticipated by Salmon [62, p. 241 and p. 263] and by Ferguson [40]. Although both Salmon and Ferguson briefly outline what we

have called the focal disk-director property and the bifocal disk property of conics, their treatments are cursory. Neither shows, as we have, that conics are completely characterized by the focal disk-director property or the bifocal disk property. In particular, neither proves that every conic has the bifocal disk property, neither relates focal disks and directors to twisted cylinders, and neither mentions infinite families of focal disks such as those described in Theorems 7.1 and 7.3.

Finally, we have pointed out geometric reasons why the twisted cylinder is more appropriate than the cone for studying conic sections. These are reenforced by analytic geometry. In Section 7.4 we found a standard cartesian equation (7.3) for a locus satisfying the focal disk-director property. In some cases, the resulting locus cannot be realized on a cone. For example, the equation $y^2 = r^2$ represents two parallel lines. Also, flipped hyperbolas cannot appear on the same cone, but they appear as sections of one twisted cylinder. Thus, the twisted cylinder is a more natural platform than the cone for investigating conics.

Interactive animation to form various twisted cylinders is available at the web site

http://www.its.caltech.edu/~mamikon/TwistCyl.html.

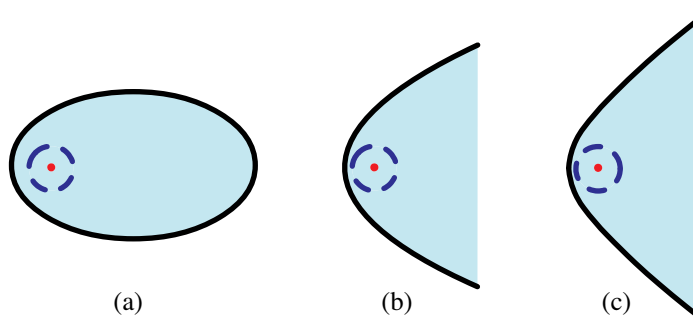
The material in this chapter was originally published in [21], except for Section 7.13, which appeared in [17].

Chapter 8

ELLIPSE TO HYPERBOLA: "WITH THIS STRING I THEE WED"

This problem can be easily solved by the methods developed in this chapter. The reader may wish to try solving it before reading the chapter.

Figure (a) shows a pool with an elliptical boundary. A pebble dropped into the pool at its left focus creates an expanding frontal circular wave centered at the focus. When the wave strikes the boundary a reflected wave is formed that always intersects the frontal wave at the boundary of the pool. A similar situation occurs in (b) with a parabolic boundary, and in (c) with a hyperbolic boundary.



Determine the shape of the reflected wave in the three cases.

Hint: All three curves, ellipse, parabola, and hyperbola, have a common reflection property that plays a role in solving this problem. For example, light rays emanating from one focus are reflected along a line passing through the other focus. For the parabola, the second focus is taken to be at infinity and the reflected rays are parallel to the axis of the parabola.

CONTENTS

8

A string mechanism is introduced that traces both elliptic and hyperbolic arcs having the same foci. This suggests replacing each focus by a focal circle centered at the focus, a simple step that leads to new characteristic properties of central conics that also extend to the parabola.

The classical description of an ellipse and hyperbola as the locus of points whose sum or absolute difference of focal distances is constant is generalized to a common bifocal property, in which the sum or absolute difference of the shortest distances to the focal circles is constant. Surprisingly, each of the sum or difference can be constant on both the ellipse and hyperbola. When the radius of one focal circle is infinite, the bifocal property becomes a new property of the parabola.

We also introduce special focal circles, called circular directrices, which provide equidistant properties for central conics analogous to the classical focus-directrix property of the parabola. Those familiar with paper-folding activities for constructing an ellipse or hyperbola using a circle as a guide, will be pleased to learn that the guiding circle is, in fact, a circular directrix. Finally, an extended eccentricity property of conics is introduced by replacing the focus by a fixed circle and the directrix by an arbitrary coplanar line.

8.1 STRING CONSTRUCTION FOR BOTH ELLIPSE AND HYPERBOLA

The title of this chapter was inspired by our modification of the well-known string construction for the ellipse. In Figure 8.1a, a piece of string joins two fixed points (the foci of the ellipse), and the string is kept taut by a moving pencil that traces the ellipse. The bifocal property of the ellipse states that the sum of distances from pencil to foci is the constant length of the string.

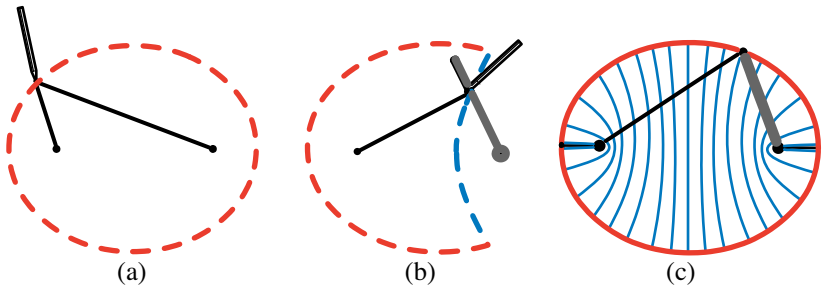


Figure 8.1: (a) String construction for the ellipse. (b) If part of the string is inside a tube of fixed length, the pencil traces a confocal hyperbolic arc. (c) If the length of the tube is varied, the pencil traces arcs of all confocal hyperbolas.

The same string fastened to the same points can also be used to trace a hyperbolic arc with the same foci. How is this possible? The bifocal property of the hyperbola states that the *difference* of distances (longer minus shorter) from any point on the hyperbola to the foci is constant. Nevertheless, a slight modification of the string construction for the ellipse shows how to do it.

The points of intersection of an ellipse with the line through its foci are called its vertices. Take a thin rigid tube shorter than the string but longer than the distance from a focus to the nearest vertex. Pass part of the string through the tube and fasten the ends of the string to the foci as before. One end of the tube pivots at a focus, like one hand of a clock. The free end traces a circle that plays a crucial role in this chapter. A pencil keeps the string taut by pushing it outward or inward in the radial direction of the tube. If it pushes outward, the tube plays no role and the pencil traces part of the ellipse as in Figure 8.1a. But if it pushes inward along the edge of the tube, it traces a portion of a confocal hyperbola lying inside the ellipse, as in Figure 8.1b. By varying the length of the tube you can draw a family of confocal hyperbolic arcs (Figure 8.1c). Because the arcs are confocal with the ellipse, they intersect it orthogonally. One of the arcs so constructed is the perpendicular bisector of the segment joining the foci.

To verify that the modified string construction works as claimed, let c denote the constant length of the string, and assume the tube has length $R < c/2$ and rotates about the focus F_2 as in Figure 8.2. The string passes through the tube, and its ends are attached to the foci F_1 and F_2. Now suppose the pencil at point P pushes the string inward toward F_2 as in Figure 8.2. The portion of the string inside the tube has length R, and the portion along the outside edge of the tube has length $R - PF_2$. Because the total length of the string is c, we have $c = R + (R - PF_2) + PF_1 = 2R - PF_2 + PF_1$. Therefore $PF_1 - PF_2 = c - 2R =$ constant, hence P lies on a hyperbola with foci F_1 and F_2, as asserted.

Contents of this chapter.

The string mechanism that weds the ellipse and hyperbola leads in a natural way to a generalization of the classical bifocal property, in which each focus is replaced

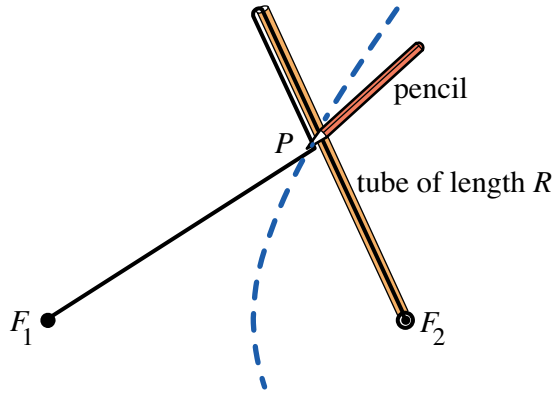


Figure 8.2: Justification of the new string construction for the hyperbola.

by a circle, called a *focal circle*, centered at that focus. Focal circles are introduced in Section 8.2, which also extends the string construction by using two tubes, each pivoted at a focus. The free end of each tube traces a focal circle. Section 8.3 provides two surprising locus properties relating the sum and difference of distances to the focal circles (Theorem 8.1) . Special pairs of focal circles, called *circular directrices*, are also introduced in Section 8.3. Those familiar with paper-folding activities for constructing an ellipse or hyperbola using a circle as a guide, will be pleased to learn that the guiding circle is, in fact, a circular directrix. Section 8.4 proves an extended bifocal property for the ellipse and hyperbola, a converse to Theorem 8.1, and shows that focal circles lead to a simple treatment of a problem of Apollonius.

Although a parabola has only one focus, Section 8.5 shows that the extended bifocal properties of the ellipse and hyperbola can be transferred to a parabola by moving one focus to ∞. In the limit, a circular directrix centered at the moving focus becomes the classical directrix of the parabola. An application is also given to a pursuit problem involving conics. Section 8.6 uses a focal circle to extend the eccentricity property common to all conics.

Focal circles should not be confused with focal disks of Chapter 7, which described conics in terms of sums and differences of tangent lengths to the focal disks.

8.2 FOCAL CIRCLES FOR ELLIPSE AND HYPERBOLA

Start with two distinct points F_1 and F_2 that will serve as foci for an ellipse or a hyperbola. Draw two coplanar circles C_1 and C_2, which we call *focal circles*, centered at the foci with respective radii $R_1 \geq 0$ and $R_2 \geq 0$. The circles may or may not intersect, and one of them might lie inside the other.

Figure 8.3 shows two intersecting focal circles that divide the plane into four regions: region 1 inside C_1 and outside C_2, region 2 inside C_2 and outside C_1, region 3 outside both C_1 and C_2, and region 4 inside both C_1 and C_2. In some cases, one of regions 1, 2, or 4 may be empty.

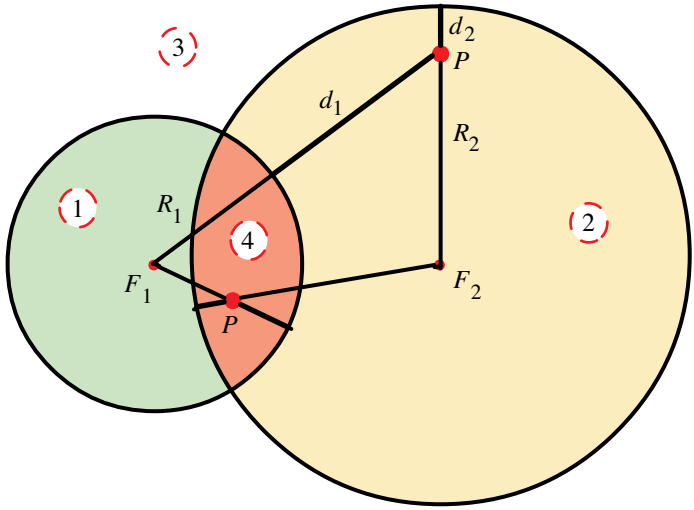


Figure 8.3: Two focal circles that divide the plane into four regions.

String mechanism involving two tubes, one at each focus.

Focal circles provide a string mechanism for tracing both elliptic and hyperbolic arcs, illustrated in Figure 8.4. Join the foci with a string of constant length $c \geq R_1 + R_2$ that passes through two thin tubes, of lengths R_1 and R_2, which pivot around the respective foci. Then $c = R_1 + R_2 + d$, where d is the length of the portion of the string outside the tubes. Note that d is constant if the sum of the radii is constant. This mechanism generalizes that in Figure 8.2, which occurs when $R_1 = 0$. An alternative but equivalent mechanism uses two clock hands of lengths R_1 and R_2 with a string of length d joining their free ends. A new feature, not needed in Figure 8.2, is the introduction of a ring that insures that the pencil keeps the string taut at the intersection of the radial directions. The four diagrams in Figure 8.4 illustrate how the mechanism works in each of the regions 3, 2, 4, and 1 of Figure 8.3.

Figure 8.5a shows a curvilinear trapezoid and its mirror image, each of which can be traced by the string mechanism in one continuous motion through all four regions in Figure 8.3. The upper trapezoid has two lower vertices on the boundary of region 4, and two upper vertices on the boundary of region 3. Place the pencil at the lower right vertex on circle C_1, moving it through region 4 to the lower left vertex on circle C_2. As we show later, this traces an arc of an ellipse (the lower edge of the trapezoid). Now continue the motion in region 1 to trace a hyperbolic arc (the left edge of the trapezoid), and then in region 3 to trace another elliptical arc (the upper edge of the trapezoid). Finally, return to the starting point by tracing another hyperbolic arc in region 2 (the right edge of the trapezoid). Theorem 8.1a will show that the length d of the portion of the string outside the tubes is the same constant on each edge of the trapezoid. By changing the value of d we obtain an entire family of trapezoids, as depicted in Figure 8.5b. As d shrinks to 0 the trapezoid becomes a point of intersection of the focal circles.

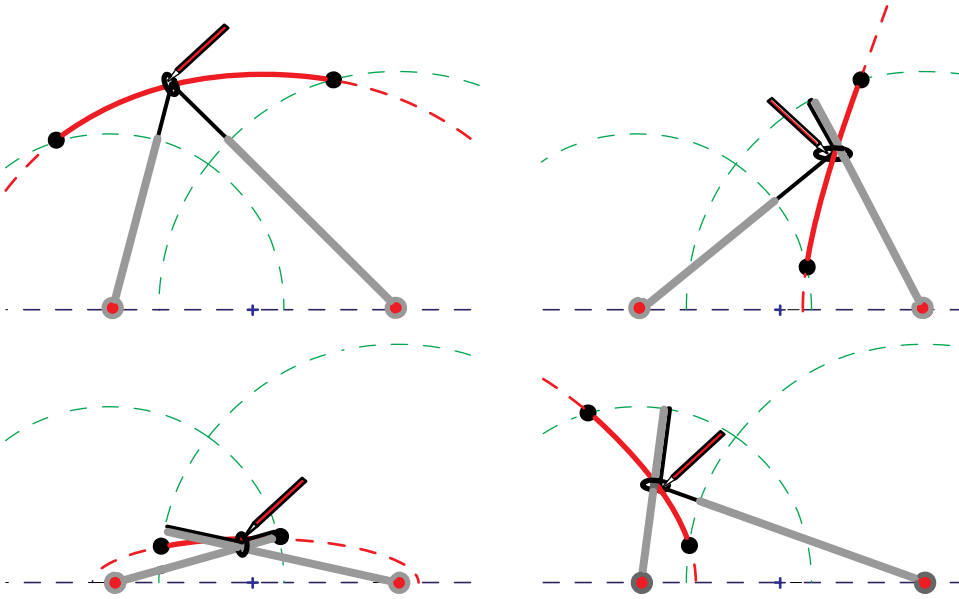


Figure 8.4: String mechanism involving two tubes. A ring keeps the string taut at the intersection of the radial directions.

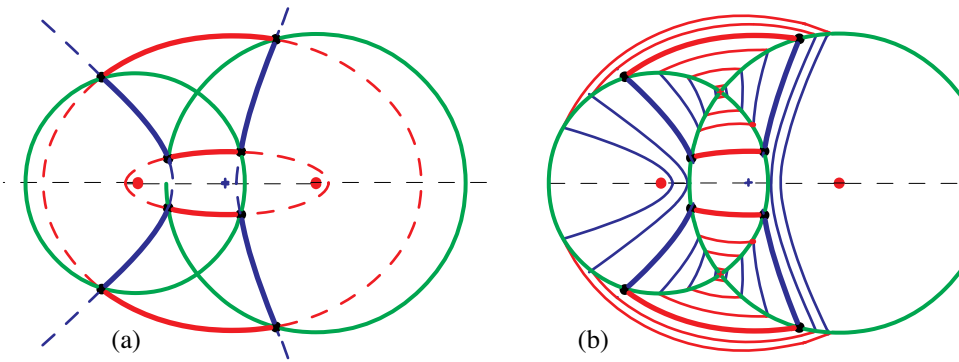


Figure 8.5: (a) A curvilinear trapezoid and its mirror image, each traced by one continuous motion of the two-tube string mechanism. (b) A family of trapezoids obtained by varying the length of the portion of the string outside the tubes.

8.3 TWO LOCUS PROPERTIES RELATING THE ELLIPSE AND HYPERBOLA

This section introduces two new and surprising locus properties relating the ellipse and hyperbola. Refer to the focal circles in Figure 8.3. Choose a point P in the plane of the circles, and let f_1 be the distance from P to F_1, and f_2 the distance from P to F_2. Also, let d_1, d_2 be the respective shortest distances from P to focal circles C_1 and C_2, each measured radially, so that $d = d_1 + d_2$ is the length of the portion of the string outside the tubes in the string mechanism. Figure 8.3 shows two choices of P, one in region 2, the other in region 4. We note that the following relations hold in Figure 8.3:

$$\text{In region 1, } d_1 = R_1 - f_1 \text{ and } d_2 = f_2 - R_2. \tag{8.1}$$

$$\text{In region 2, } d_1 = f_1 - R_1 \text{ and } d_2 = R_2 - f_2. \tag{8.2}$$

$$\text{In region 3, } d_1 = f_1 - R_1 \text{ and } d_2 = f_2 - R_2. \tag{8.3}$$

$$\text{In region 4, } d_1 = R_1 - f_1 \text{ and } d_2 = R_2 - f_2. \tag{8.4}$$

By adding d_1 and d_2 in each region we obtain:

Lemma 8.1. (1) *If P is in region* 1, *then* $d_1 + d_2 = (f_2 - f_1) - (R_2 - R_1)$.
(2) *If P is in region* 2, *then* $d_1 + d_2 = (f_1 - f_2) - (R_1 - R_2)$.
(3) *If P is in region* 3, *then* $d_1 + d_2 = (f_1 + f_2) - (R_1 + R_2)$.
(4) *If P is in region* 4, *then* $d_1 + d_2 = (R_1 + R_2) - (f_1 + f_2)$.

When $d_1 + d_2$ is constant, Lemma 8.1 reveals the following information about the curves traced by the string mechanism:

In region 1, $f_2 - f_1$ is constant and P traces part of a hyperbola with foci F_1 and F_2. In Figure 8.5a, this part is shown as two solid arcs on the left branch of this hyperbola.

In region 2, $f_1 - f_2$ is a different constant and P traces part of a different hyperbola with the same foci. In Figure 8.5a, this part is shown as two solid arcs on the right branch of the second hyperbola.

In region 3, $f_1 + f_2 = R_1 + R_2 + d_1 + d_2 = c$, the length of the string, and P traces part of an ellipse, shown in Figure 8.5a as two solid elliptical arcs.

In region 4, the constant focal sum $f_1 + f_2$ differs from that in region 3, and P traces two solid arcs of the smaller ellipse shown in Figure 8.5a.

By subtracting distances d_1 and d_2 in each region, we obtain:

Lemma 8.2. (1) *If P is in region* 1, *then* $d_2 - d_1 = (f_1 + f_2) - (R_1 + R_2)$.
(2) *If P is in region* 2, *then* $d_1 - d_2 = (f_1 + f_2) - (R_1 + R_2)$.
(3) *If P is in region* 3, *then* $d_1 - d_2 = (f_1 - f_2) - (R_1 - R_2)$.
(4) *If P is in region* 4, *then* $d_2 - d_1 = (f_1 - f_2) - (R_1 - R_2)$.

When $|d_1 - d_2|$ is constant, Lemma 8.2 reveals the following information about the curves traced by the string mechanism:

In regions 1 and 2, the focal sum $f_1 + f_2 = R_1 + R_2 + |d_1 - d_2|$ is constant, so P traces an elliptical arc. Each of these arcs, shown dashed in Figure 8.5a, is a continuation of a corresponding solid elliptical arc in Figure 8.5a, because their focal sums agree at those points where the arcs intersect the focal circles.

A similar analysis shows that each dashed hyperbolic arc in regions 3 and 4 of Figure 8.5a is a continuation of a corresponding solid hyperbolic arc in regions 1 and 2. Thus, from Lemmas 8.1 and 8.2 we deduce:

Theorem 8.1. (a) *The locus of points P such that $d_1 + d_2$ is constant is part of an ellipse or a hyperbola.*
(b) *The locus of points P such that $|d_1 - d_2|$ is constant is part of an ellipse or a hyperbola.*

Proof. Part (a) follows from Lemma 8.1, and part (b) from Lemma 8.2.

Theorem 8.1 uncovers the surprising fact that use of focal circles allows each of $d_1 + d_2$ and $|d_1 - d_2|$ to be constant on both the ellipse and hyperbola.

Circular directrices for the ellipse and hyperbola.

Unlike central conics (ellipse and hyperbola), which have two foci, a parabola has only one focus F. It can be described as the locus of a point P that moves in a plane with its focal distance PF always equal to its distance PD from a fixed line D, called its directrix. Now we introduce an analogous equidistant property for central conics, using two special focal circles with the property that *from each point of the central conic, the shortest distances to the two focal circles are equal.* If such focal circles exist, we call them *circular directrices.* Now we will show that Lemma 8.2 implies that they do exist and tells us how to determine them.

A point P is equidistant (equal shortest distances) from the two focal circles if, and only if, $d_1 = d_2$. According to Lemma 8.2, this happens in regions 1 and 2 of Figure 8.3 when $R_1 + R_2 = f_1 + f_2$, and the same occurs in regions 3 and 4 when $R_1 - R_2 = f_1 - f_2$.

On an ellipse, the constant sum $f_1 + f_2$ represents the length of the major axis of the ellipse, while on a hyperbola the constant difference $|f_1 - f_2|$ represents the length of the transverse axis. This gives us the following:

Description of circular directrices. *Any two focal circles with sum of radii $R_1 + R_2 = f_1 + f_2$ serve as a pair of circular directrices for the ellipse; any two focal circles with $|R_1 - R_2| = |f_1 - f_2|$ serve as a pair of circular directrices for the hyperbola.*

Each central conic has infinitely many pairs of circular directrices. Figure 8.6a shows three pairs of circular directrices for an ellipse. In (a), $R_1 = 0$, C_1 becomes the focus F_1, and $R_2 = f_1 + f_2$, so the entire ellipse lies inside the focal circle C_2, hence in region 2. Each point on the ellipse is equidistant from focus F_1 and from this particular focal circle C_2, as depicted in Figure 8.6a. This circular directrix C_2 is used in standard paper-folding constructions for the ellipse. It was also used by Feynman [42; p. 152] in his geometric treatment of Kepler's laws of planetary motion.

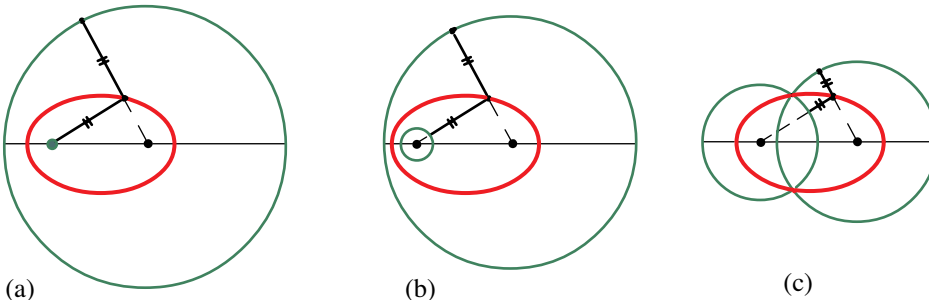


Figure 8.6: An ellipse and three pairs of its circular directrices. In (a), $R_1 = 0$ and $R_2 = f_1 + f_2$. In (b) and (c), $R_1 > 0$ and $R_2 = f_1 + f_2 - R_1$. For each P on the ellipse, the shortest distances d_1 and d_2 to the focal circles are equal.

In Figures 8.6b and 8.6c, both circular directrices have positive radii, but the sum of the radii is constant, so an increase in R_1 results in a corresponding decrease in R_2. In Figure 8.6b, the entire ellipse is in region 2, but in Figure 8.6c part of the ellipse is in region 2 and the remaining part in region 1.

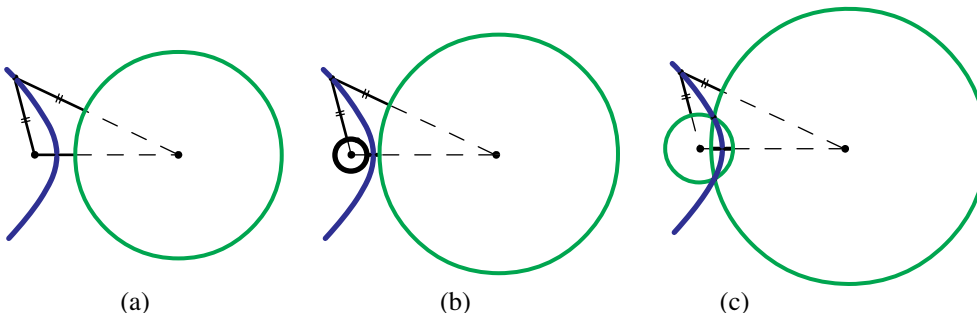


Figure 8.7: One branch of a hyperbola and three pairs of its circular directrices. In (a), $R_1 = 0$, $R_2 = f_2 - f_1$. In (b) and (c), $R_1 > 0$, $R_2 = f_2 - f_1 + R_1$. For each P on the hyperbola, the shortest distances d_1 and d_2 to the focal circles are equal.

Figure 8.7 shows three pairs of circular directrices for one branch of a hyperbola. In Figure 8.7a, $R_1 = 0$ and $R_2 = f_2 - f_1$. This circular directrix is used in standard paper-folding constructions for the hyperbola. In Figures 8.7b and 8.7c, $R_1 > 0$ and $R_2 = f_2 - f_1 + R_1$, so an increase in R_1 results in a corresponding increase in R_2.

One may very well ask *"What about the other branch of the hyperbola"?*

Figure 8.8 shows both branches and reveals something new. From any point P there are two distances to each focal circle, the *shortest* distances, which we have denoted by d_1 and d_2, and the *longest* distances which we denote by D_1 and D_2. The difference $D_i - d_i$ is $2R_i$, the diameter of focal circle C_i. The longest distance D_2 plays a role on the second branch. Figure 8.8a shows both branches of the

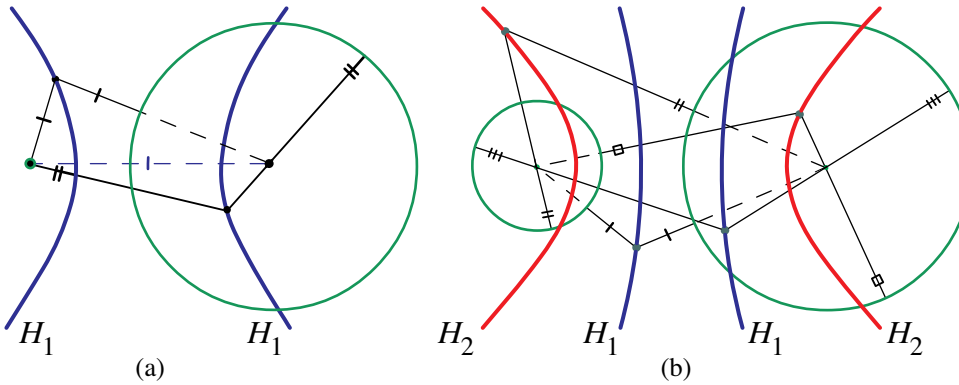


Figure 8.8: (a) Both branches of hyperbola in Figure 8.7a. For each P on the right branch the shortest distance d_1 is equal to the longest distance D_2. (b) Two confocal hyperbolas H_1 and H_2 with transverse axes of different lengths.

hyperbola in Figure 8.7a, labeled as H_1. Here we have $d_1 = D_2$ on the second branch. In other words, the shortest distance to C_1 is equal to the longest distance to C_2. In this case, $C_1 = F_1$ because $R_1 = 0$.

But when $R_1 > 0$ and $R_2 = f_2 - f_1 + R_1$, a new phenomenon occurs. A second hyperbola comes into play with the same foci but with a different transverse axis, as shown in Figure 8.8b. Let H_1 denote the hyperbola with the shorter transverse axis, and H_2 the one with the longer. Each point on the left branch of H_1 has $d_1 = d_2$, as in Figure 8.7, but each point on the right branch of H_1 has $D_1 = D_2$. On the left branch of H_2 we have $D_1 = d_2$, and on the right branch of H_2 we have $d_1 = D_2$, as indicated by the tick marks in Figure 8.8b. Thus, the circular directrices play a deeper role than was revealed in Figure 8.6. This is illustrated further in Figure

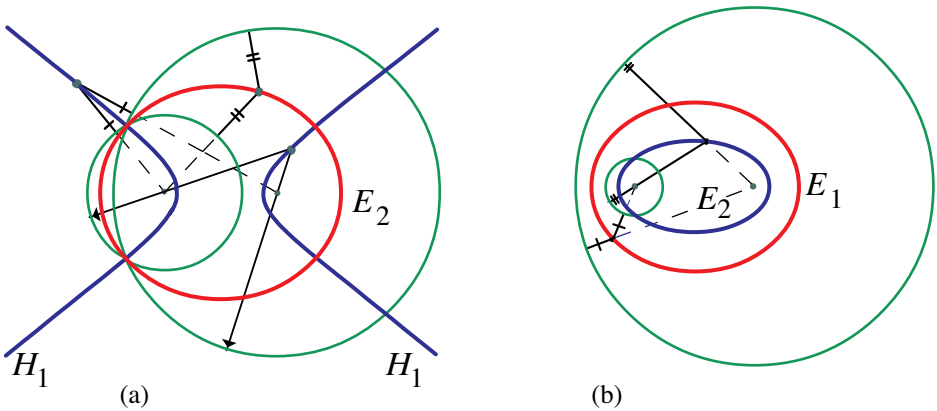


Figure 8.9: (a) Hyperbola H_2 in Figure 8.8b is replaced by ellipse E_2. (b) Hyperbolas H_1 and H_2 in Figure 8.8b are replaced by ellipses E_1 and E_2, respectively.

8.9a, which can be thought of as a continuation of Figure 8.8b. As R_2 increases, the asymptotes of H_2 become more and more horizontal until R_2 reaches a critical value for which H_2 degenerates to a pair of rays emanating from the foci.

For points on the degenerate hyperbola, $|f_2 - f_1|$ is the distance between the foci, which is also $f_1 + f_2$, the sum of focal distances from points on the line segment joining the foci. This segment is a degenerate ellipse. As R_2 increases beyond the critical value and the circular directrices C_1 and C_2 intersect as shown in Figure 8.9a, hyperbola H_2 in Figure 8.8b is replaced by a confocal ellipse E_2 on which $d_1 = d_2$. On the left branch of H_1 we have $d_1 = d_2$, and on its right branch we have $D_1 = D_2$, as in Figure 8.8b. As R_2 increases further, so that C_2 contains C_1 in its interior, as in Figure 8.9b, H_1 also degenerates and is replaced by a second confocal ellipse E_1 on which $d_1 = d_2$.

In what follows, results are stated in terms of the shortest distance d_i to the focal circle C_i. Any problem involving the longest distance D_i can be reduced by the substitution $D_i = d_i + 2R_i$ to an equivalent problem involving the shortest distance d_i.

8.4 EXTENDED BIFOCAL PROPERTY: ELLIPSE AND HYPERBOLA

The next theorem provides an extended bifocal property that has the same form for the ellipse and the hyperbola. It is stated in terms of shortest distances d_1, d_2 to the focal circles. The sum of focal distances $f_1 + f_2$ from a point on an ellipse to its foci is a constant, the length of the major axis, which we denote by A. On a hyperbola, the difference $|f_1 - f_2|$ of focal distances is another constant, the length of the transverse axis, which we denote by B. On the left branch (enclosing F_1) we have $f_2 - f_1 = B$, and on the other branch we have $f_1 - f_2 = B$.

Theorem 8.2. (a) *Given an ellipse with major axis of length A, and two focal circles C_1, C_2 of radii R_1, R_2, let $d = R_1 + R_2 - A$. Then each point on the ellipse satisfies one of*

$$d_1 + d_2 = |d| \tag{8.5}$$

$$|d_1 - d_2| = |d|. \tag{8.6}$$

(b) *Given a hyperbola with transverse axis of length B, and given the same focal circles as in* (a)*, let $d' = B - (R_1 + R_2)$. Then each point on the left branch satisfies one of*

$$d_1 + d_2 = |d'| + 2R_1 \tag{8.7}$$

$$|d_1 - d_2| = |d'| + 2R_1. \tag{8.8}$$

Each point on the right branch satisfies one of

$$d_1 + d_2 = |d'| + 2R_2 \tag{8.9}$$

$$|d_1 - d_2| = |d'| + 2R_2. \tag{8.10}$$

If $R_1 = R_2$, then on both branches we have $d_1 + d_2 = B$ or $|d_1 - d_2| = B$.

Proof of (a). We consider cases, depending on the sign of d.

Case 1: $d \le 0$, so $R_1 + R_2 \le A$.

If both focal circles lie inside the ellipse, then $d_1 = f_1 - R_1$ and $d_2 = f_2 - R_2$, hence $d_1 + d_2 = f_1 + f_2 - (R_1 + R_2) = A - (R_1 + R_2) = -d = |d|$, so (8.5) holds on the entire ellipse.

If the ellipse intersects a focal circle, say C_1, then at the point of intersection we have $d_1 = 0, f_1 = R_1, d_2 = f_2 - R_2 = A - f_1 - R_2 = A - (R_1 + R_2) = -d = |d|$, so both (8.5) and (8.6) hold at this point. But Lemma 8.1(3) shows that for this value of d, (8.5) holds for every point of the ellipse in regions 1 or 2. A similar argument works if the ellipse intersects focal circle C_2. This proves (a) in Case 1.

Case 2: $d > 0$, so $R_1 + R_2 > A$.

Now the focal circles intersect each other and also intersect the ellipse. At a point where C_1 intersects the ellipse we have $d_1 = 0, f_1 = R_1, d_2 = f_2 - R_2 = A - f_1 - R_2 = A - (R_1 + R_2) = -d$, so $d_1 - d_2 = d = |d|$ at the point of intersection. By Lemma 8.2(1), $d_1 - d_2 = d$ for every point of the ellipse in region 1, and $d_2 - d_1 = d$ in region 2. Also, Lemma 8.1(4) shows that $d_1 + d_2 = d$ for every point of the ellipse in region 4. This proves (a) in Case 2.

Proof of (b). On a hyperbola $|f_1 - f_2|$ is constant, so $B = |f_1 - f_2|$. Again we consider two cases, depending on the relation between B and $R_1 + R_2$.

Case 1: $B \ge R_1 + R_2$.

In this case the focal circles do not intersect, and at most one of them can intersect the hyperbola. If neither focal circle intersects the hyperbola, then $f_1 = R_1 + d_1$ and $f_2 = R_2 + d_2$, so $d_1 - d_2 = f_1 - f_2 + R_2 - R_1$, which is the same as $d_2 - d_1 = f_2 - f_1 + R_1 - R_2$. On the left branch, $f_1 < f_2$, so $f_2 - f_1 = B$ and $d_2 - d_1 = B + R_1 - R_2 = B - (R_1 + R_2) + 2R_1$, so (8.8) is satisfied everywhere on this branch.

On the right branch, $f_2 < f_1$, so $f_1 - f_2 = B$ and $d_1 - d_2 = B + R_2 - R_1 = B - (R_1 + R_2) + 2R_2$, so (8.10) is satisfied everywhere on this branch.

Now suppose that one focal circle, say C_1, intersects the hyperbola. At a point of intersection we have $d_1 = 0, R_1 = f_1$ so $d_2 - d_1 = f_2 - R_2 = f_2 - f_1 + R_1 - R_2 = B + R_1 - R_2 = B - (R_1 + R_2) + 2R_1$. By Lemma 8.2(3), $d_2 - d_1$ has the same value everywhere in region 3, so (8.8) holds everywhere in region 3. Now by Lemma 8.1(1), in region 1 we have $d_1 + d_2 = f_2 - f_1 - R_2 + R_1 = B - (R_1 + R_2) + 2R_1$, so (8.7) holds everywhere in region 1. Therefore either (8.7) or (8.8) holds on the left branch, whereas (8.10) holds everywhere on the right branch. If, however, C_2 intersects the hyperbola, then the same type of argument shows that either (8.9) or (8.10) holds on the right branch, and (8.8) holds everywhere on the left branch. This proves (b) in Case 1.

Case 2: $B < R_1 + R_2$.

In this case the focal circles overlap and each intersects the hyperbola. The same type of argument used for Case 1 shows that, on the left branch, (8.7) holds in regions 3 and 4, and (8.8) holds in region 1. Similarly, on the right branch, (8.9) holds in regions 3 and 4, and (8.10) holds in region 2.

Application to Apollonius's kissing circles problem.

Focal circles for central conics can be used to give a simple treatment of a classical three-circle problem attributed to Apollonius [47; p. 182]. First we consider the following:

Two-circle locus problem. Choose two distinct circles C_1 and C_2 in the same plane with centers at points F_1 and F_2. The circles may or may not intersect, and one of them might lie inside the other. We are interested in finding a third circle C tangent to both C_1 and C_2.

Solution. There are, of course, infinitely many such circles C. By regarding C_1 and C_2 as focal circles, it is easy to show that the locus of all their centers is a central conic with foci at F_1 and F_2. We omit details except to suggest consideration of the following cases:

(a) Nonintersecting C_1 and C_2, neither inside the other. In this case the locus is a hyperbola if (a_1) both C_1 and C_2 are inside C, (a_2) both are outside C, and (a_3) one is inside C and the other is outside C.

(b) Nonintersecting C_1 and C_2, with C_1 inside C_2. In this case the locus is an ellipse. There are two subcases: (b_1) C inside C_2 and outside C_1, and (b_2) C_1 inside C inside C_2.

(c) Intersecting C_1 and C_2. In this case the locus is (c_1) an ellipse if C is inside one of C_1 or C_2 and outside the other, or (c_2) a hyperbola if C is outside both C_1 and C_2, or inside both C_1 and C_2.

We turn now to the three-circle problem of Apollonius:

Apollonius problem. *Choose three coplanar circles C_1, C_2, C_3 with respective centers at F_1, F_2, F_3. Find a fourth circle C tangent to all of the given circles.*

If C exists, it is tangent to each pair of the three circles. By considering one pair of circles at a time, say C_1 and C_2, our two-circle locus problem tells us that the center of C lies on a central conic with foci F_1, F_2. Similarly, it lies on a central conic with foci F_1, F_3, and on a central conic with foci F_2, F_3. The point of intersection of these central conics is the center of C, which is equidistant to each of the C_i.

8.5 BIFOCAL PROPERTIES TRANSFERRED TO THE PARABOLA

The extended bifocal properties of the central conics were obtained by replacing each focus by a focal circle. Although the parabola has only one focus, we can transfer the extended bifocal properties to the parabola by keeping one focal circle fixed and moving the second focus to ∞, allowing the radius of the second focal circle to increase without bound. The second focal circle now becomes a line perpendicular to the focal axis, which we call a *floating focal line*. The central conic becomes a parabola whose focus is the center of the fixed focal circle, and whose directrix is parallel to the floating focal line. As expected, the bifocal properties of the central conics can be transferred to the focal circle and the floating focal line.

This process is consistent with the geometric definition of conics as sections of a cone. Recall that a plane cutting one nappe of a right circular cone produces an ellipse. As the plane is tilted to become nearly parallel to a generator of the

cone, the ellipse becomes more elongated, and when the cutting plane is parallel to a generator the intersection becomes a parabola. Tilt the plane even further so it cuts both nappes, and the intersection is a hyperbola. Thus, as a section of a cone, the parabola is a transition between the ellipse and hyperbola, so it is not surprising that properties of the parabola can be obtained as limiting cases of those of a central conic.

First we introduce the parabolic version of circular directrices. For central conics, circular directrices occur in pairs, each a special focal circle. In the parabolic version, one of the circular directrices is replaced by a limiting line called a floating directrix.

Pairs of circular and floating directrices for the parabola.

A parabola has only one focus F, and is the locus of a point P that moves in a plane with its focal distance PF always equal to its distance PD from a fixed line D, called its directrix (Figure 8.10a). We call D the *linear directrix* of the parabola, to distinguish it from *circular directrices*, which we define as follows. A line L parallel to directrix D we call a *floating focal line*. In Figures 8.10b and 8.10c, L is between F and D. Let R denote the distance between L and D. Then the focal circle $C(R)$ of radius R and center F is called a *circular directrix* for the parabola, *relative to* L. In this context, L is also called a *floating directrix* corresponding to the circular directrix. This terminology was chosen because for every point P on the parabola we have $d_C = d_L$, where d_C is the shortest distance from P to focal circle $C(R)$, and d_L is the distance from P to L. The common distance is $PF - R$, as shown in Figures 8.10b and 8.10c. *When the radius of C is zero, circular directrix C becomes the focus of the parabola, and floating directrix L becomes its classical directrix D,* as in Figure 8.10a.

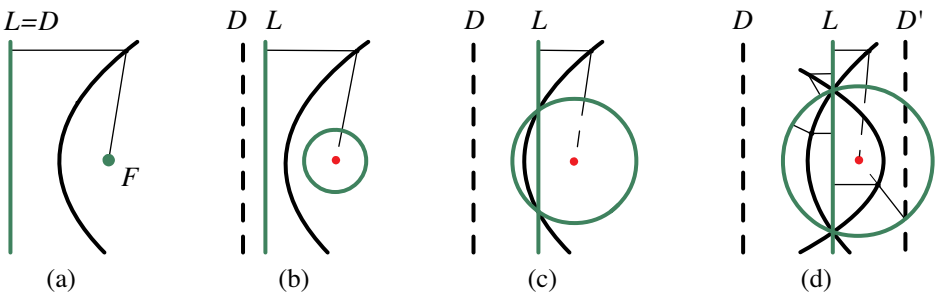


Figure 8.10: (a) Focus and directrix of a parabola. In (b)-(d), circular directrix and floating directrix. In (d), two intersecting confocal parabolas with linear directrices D and D', but with the same circular directrix and floating directrix.

As expected, each circular directrix of a parabola can be obtained as the limiting case of a circular directrix of a central conic by sending one of the foci to ∞. To illustrate, begin with an ellipse and two circular directrices C_1, C_2, as in Figure 8.6b or 8.6c, where $d_1 = d_2$ for each point on the ellipse. Let Q be the point where C_2 intersects the focal axis. Keep F_1, R_1, and Q fixed, and move F_2 along the focal axis arbitrarily far away, so that $R_2 \to \infty$. Then, the limiting circle C_2 becomes a line

L through Q perpendicular to the focal axis. The radial distance d_2 becomes d_L, the distance from P to L, and the ellipse becomes a limit curve with the property that $d_1 = d_L$ at each of its points. This limit curve is, in fact, a parabola with focus F_1 and linear directrix D, whose distance from L is R_1, because each of its points is equidistant from F_1 and D. The circular directrix C_1 for the ellipse is now a circular directrix for the parabola with L as its floating directrix. The parabola opens to the right, as in Figures 8.10b and 8.10c. The choice of Q determines the position of the floating directrix L.

We can arrive at the same circular directrix and the same floating directrix L by starting with the left branch of the hyperbola shown in Figure 8.7, keeping F_1, R_1, and Q fixed as before, and letting $R_2 \to \infty$. If Q is between F_1 and the vertex of the left branch, as in Figure 8.7c, the limit curve is a parabola that opens to the left and intersects the first parabola, as shown by the example in Figure 8.10d, with its linear directrix D' parallel to L. Both parabolas intersect the floating directrix L and the circular directrix at the same points.

For a fixed focus F_1, the resulting parabola depends on two parameters, R_1 and Q. Figure 8.11 shows what happens when R_1 is fixed and the initial point Q moves to the right, starting from a point to the left of circular directrix C_1. The floating directrix L, which always passes through Q, moves as shown in Figure 8.11. In Figures 8.11a and 8.11b, the parabola opens to the right, and its linear directrix D is at a distance R_1 to the left of L, with $d_1 = d_L$ at each point P of the parabola. But when L becomes tangent to the circular directrix, as in Figure 8.11c, a second parabola enters the scene. For a point P on the portion of the focal axis with right endpoint at the focus, we have $d_1 = d_L$, just as on the earlier parabolas. So this portion of the focal axis can be regarded as a degenerate parabola with the same focus, and with the same circular directrix. But its linear directrix is a line D' parallel to L passing through the focus.

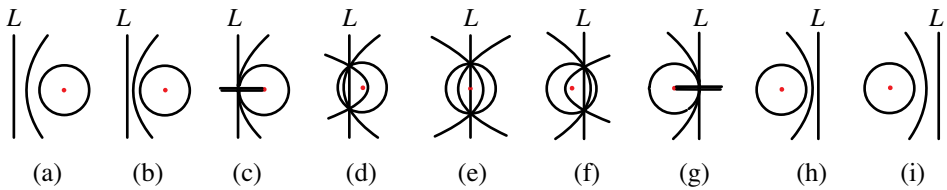


Figure 8.11: In (a) and (b), the floating directrix L is to the left of the circular directrix C_1. In (c) through (g), it intersects C_1, and in (h) and (i) it is to the right of C_1. The parabolas in (f) through (i) are mirror images of those in (d) through (a).

As L continues moving to the right, *two* parabolas appear with the *same circular directrix*, as in Figures 8.11d, e, and f. They have a common focus F_1 and a common floating directrix L. One opens to the right, the other to the left, and they intersect on L. Their linear directrices D and D' are parallel, each at distance R_1 from L. When the floating directrix L becomes tangent to the other side of the circular

directrix, as in Figure 8.11g, the first parabola degenerates to another ray, the portion of the focal axis with its left endpoint at the focus. In Figures 8.11h and i, L moves further to the right, and we are back to one parabola that now opens to the left. The parabolas in Figures 8.11f, g, h, i, are mirror images of those in Figures 8.11d, c, b, a, respectively.

We could keep the initial point Q fixed and allow R_1 to increase and obtain a new sequence of parabolas analogous to those in Figure 8.11 with focal circles of increasing radii. But, by scaling these circles so they have equal radii, we get a collection of parabolas resembling those that are shown in Figure 8.11.

Parabolic version of the extended bifocal properties.

The extended bifocal properties of central conics in Theorems 8.1 and 8.2 have counterparts for the parabola. They can be obtained by starting with two focal circles C_1 and C_2 of a central conic, and letting the radius of one of them, say R_2, go to ∞, keeping F_1, R_1, and Q fixed, as was done earlier. The limiting C_2 becomes a floating focal line L through Q perpendicular to the focal axis, and the limiting central conic becomes a parabola with focus F_1 and focal circle C_1. The new properties relate C_1 and L.

Figure 8.12a shows what happens to Figure 8.3 in the limiting case, and Figure 8.12b shows how the four central conics in Figure 8.5a become four parabolas.

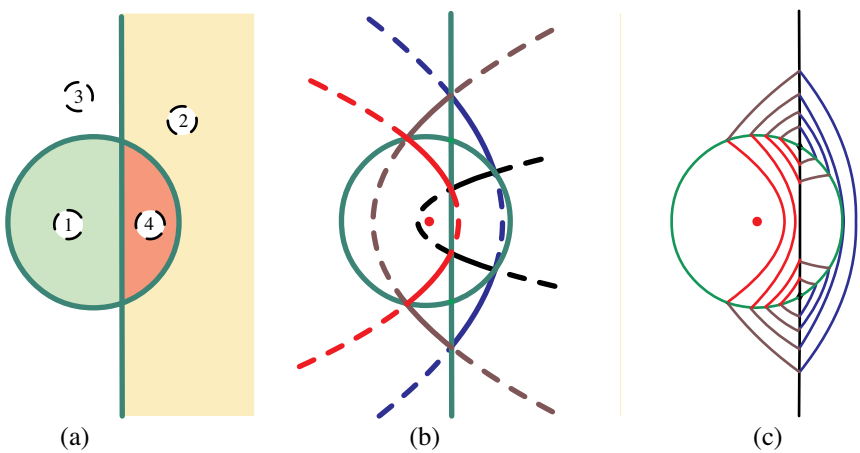


Figure 8.12: (a) Four regions formed by a focal circle and a coplanar line. (b) Limiting case of Figure 8.5a when the ellipses and hyperbolas become four parabolas. (c) A family of parabolic trapezoids obtained as the limiting case of the family in Figure 8.5b.

On the solid portions of the parabolas, the sum of distances $d_1 + d_L$ to the focal circle and the floating focal line is constant, just as on the corresponding ellipses and hyperbolas in Figure 8.5a. On the dashed portions, the absolute difference $|d_1 - d_L|$ is constant, just as on the corresponding central conics in Figure 8.5a. Figure 8.12b

also illustrates the following parabolic counterparts of Theorems 8.1 and 8.2, whose proofs are omitted.

Theorem 8.3. *Given a circle C with center at F, and a coplanar line L, if P is in the plane of C and L, let d_C and d_L denote the shortest distances from P to C and L, respectively. Then the locus of points P such that either the sum $d_C + d_L$ or the absolute difference $|d_C - d_L|$ is constant is part of a parabola with focus F and directrix parallel to L.*

Theorem 8.4. *Given a parabola with a focal circle C, and given any line L parallel to its directrix D, whose distance from D is the radius of C, then either the sum $d_C + d_L$ or the absolute difference $|d_C - d_L|$ is constant.*

String construction for the parabola.

When focal circle C has radius 0, the property that $d_C + d_L$ is constant reduces to $d_F + d_L$ is constant, and leads to a string construction for the parabola, illustrated in Figure 8.13a. One endpoint of a string of constant length is fastened to a fixed point, but the other end is attached to a small ring that slides freely along a rigid rod (a fixed line) that may or may not pass through the fixed point. The string is kept

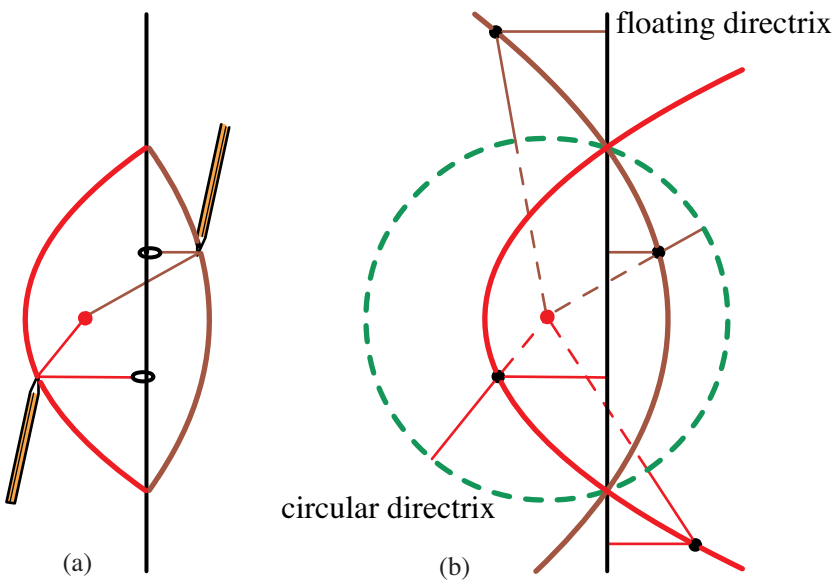


Figure 8.13: (a) String construction that gives two confocal parabolas, one on either side of the fixed line. (b) Justification of construction, using the common circular directrix of the two parabolas, with the corresponding floating directrix.

taut by a pencil that moves so that the sum of distances from the pencil to the point and to the line is the constant length of the string. The pencil traces a portion of a parabola with the fixed point as its focus. A second parabola with the same focus

can be drawn by placing the pencil on the other side of the fixed line, as indicated in Figure 8.13a. Kepler [53; p. 110] devised a similar string construction for the parabola that does not use a ring and produces only one of the two parabolas. Our construction for two confocal parabolas is justified by the diagram in Figure 8.13b, which shows the common circular directrix of the two parabolas with the fixed line as floating directrix.

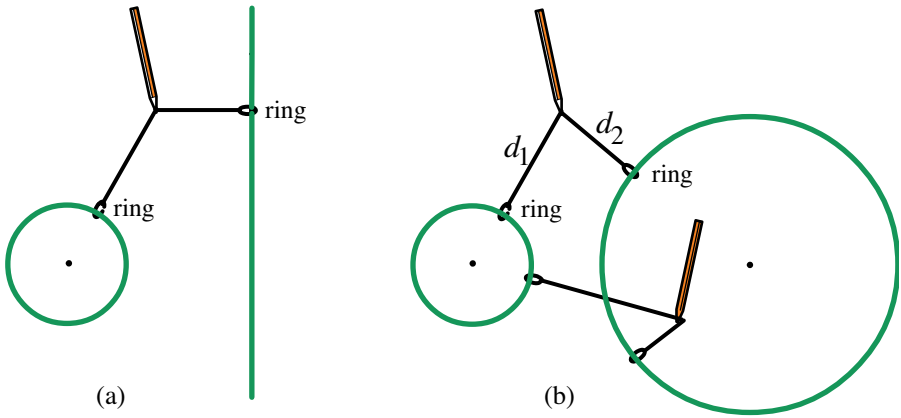


Figure 8.14: Alternative string construction (a) for parabolas, using a focal circle, and (b) for central conics, using two focal circles.

Figure 8.14a shows an equivalent form of the string construction in Figure 8.13a with a focal circle C of radius $R > 0$ centered at F. Then $d_C = d_F - R$, and the constant sum $d_F + d_L$ is replaced by $d_C + d_L = d_F + d_L - R$, another constant. The small ring that moves freely around the rigid boundary of the focal circle allows the pencil to trace the parabolas in Figure 8.13. Similarly, the two tubes and the single ring in the string mechanism of Figure 8.4 could be replaced by two small rings (illustrated by two examples in Figure 8.14b) that move freely around the two focal circles, with one end of the string attached to each ring. The pencil keeps the string taut so that the two portions of the string are in the appropriate radial directions.

Application of Theorem 8.3 to a pursuit problem.

A classical pursuit problem involves an aircraft flying at constant speed $v > 0$ from a point A toward a base F (Figure 8.15a). Because visibility is limited, an automatic pilot always aims the aircraft toward F. Ordinarily, the path would be along a straight line from A to F. However, a steady north wind with constant speed w forces the aircraft off course, so its trajectory is along a curved path that depends on the ratio of the speeds v and w. The problem is to determine the path.

If $w > v$, the craft cannot overcome the influence of the wind and moves further away from the base, approaching asymptotically the line due north from F. But if $v > w$, the aircraft overcomes the influence of the wind and returns to F along a curved path. These two solutions, which seem intuitively reasonable, can also be

verified analytically by solving a suitable differential equation. The dashed curves in Figure 8.15a indicate the qualitative nature of the solutions. The line segment from A to F shows the path when $w = 0$.

The case of interest for us is when $w = v$. In this case the solution of the

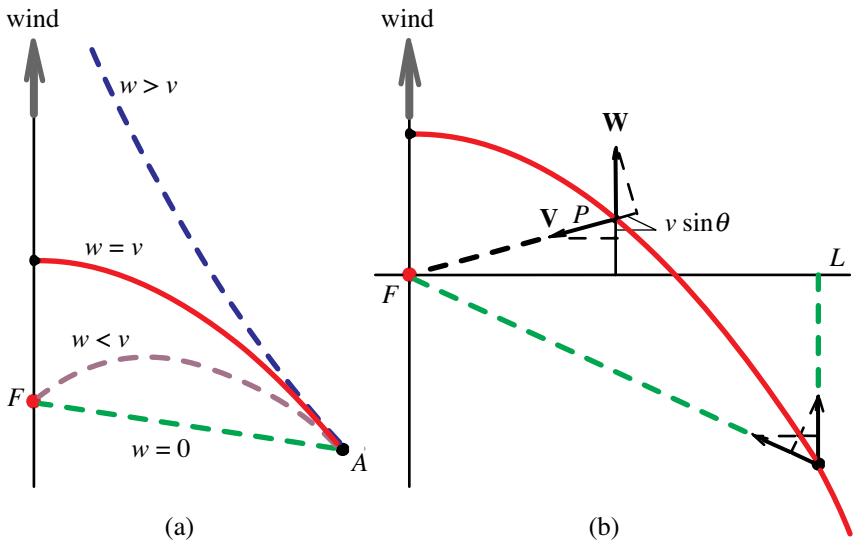


Figure 8.15: (a) Qualitative shape of trajectory depends on speeds v and w. (b) When $v = w$, the sum $d_F + d_L$ is constant above L, whereas $d_F - d_L$ is constant below L, hence the trajectory is a parabola with focus F.

differential equation is part of a parabola, shown as the solid curve in Figure 8.15a. The point F is the focus of this parabola. The aircraft moves along the parabola until it is due north of F at which point it remains stationary because the effect of its speed and that of the wind cancel each other. We shall obtain this solution by applying Theorem 8.3.

Choose a line L through F perpendicular to the wind direction, as in Figure 8.15b. We regard L as a focal line, and let F serve as a focal point. Let P denote a general point on the path of the aircraft, and let d_F and d_L denote its distances from F and L, respectively, as indicated in Figure 8.15b. The line L divides the trajectory into two parts, one above L and one below. We will show that the sum $d_F + d_L$ is constant when P is above L, and that the difference $d_F - d_L$ is the same constant when P is below L. By Theorem 8.3 with $C = F$, this will prove that the path is a parabola with focus F.

Suppose P is above L. Let θ denote the angle between L and the line joining F to P. In general, P moves along a tangent vector to the path with velocity $\mathbf{V}+\mathbf{W}$, the resultant of two vectors $\mathbf{V}$ and $\mathbf{W}$ of lengths $v = |\mathbf{V}|$ and $w = |\mathbf{W}|$. We are considering the case in which $w = v$. The vector $\mathbf{W}$, in the direction of the wind, acts to increase d_L at the time rate v. But $\mathbf{V}$ acts to decrease d_L by a component of magnitude $v \sin\theta$ opposite to $\mathbf{W}$. Hence the resultant $\mathbf{V}+\mathbf{W}$ has a component

in the direction of **W** equal to $v - v \sin\theta$, which represents the time rate of change of d_L. Similarly, the component of the resultant in the direction of **V** is $v \sin\theta - v$, which represents the time rate of change of d_F. Therefore the time rate of change of the sum $d_F + d_L$ is zero, hence $d_F + d_L$ is constant. This constant is d_F when $d_L = 0$, and is the distance from F to the point where L intersects the trajectory.

When P is below L the analysis is similar, except that both **V** and **W** act to decrease d_L so the resultant **V**+**W** has a component in the direction of **W** of magnitude $v + v \sin\theta$, whose negative is the time rate of change of d_L, and a component in the direction of **V** of the same magnitude, whose negative is the time rate of change of d_F. Therefore the time rate of change of the difference $d_F - d_L$ is zero, so $d_F - d_L$ is constant, the same constant obtained when P is above L. Thus the trajectory satisfies Theorem 8.3, so it is a parabola with focus at F.

Modified pursuit problem.

The problem can be modified so that the parabola is replaced by other conics. Specifically, suppose a wind of constant speed v blows radially outward from a point F_0 different from F. In this case one can verify, with analysis similar to that given above, that the aircraft moves along a portion of an ellipse with one focus at F_0 and the other focus at F. This application may not conform to reality, but other more realistic physical situations can be imagined that involve the same ideas. If the wind blows radially inward toward F_0, then the aircraft moves along a portion of a hyperbola.

8.6 CIRCULAR DIRECTRICES AND WAVE MOTION

A pebble dropped into a pool of water creates a frontal circular wave that expands radially along the surface. When the wave strikes the boundary of the pool, a reflected wave is formed, whose shape depends on the shape of the boundary. Now assume the boundary of the pool is an ellipse and drop the pebble at a focus. *What is the shape of the reflected wave*? We will show that it, too, is a circular arc centered at the other focus. Moreover, at each instant the focal circles containing the wave and its reflection are a pair of circular directrices for the ellipse.

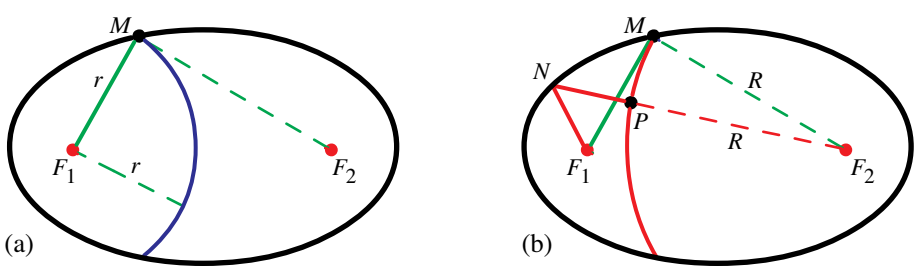


Figure 8.16: (a) Frontal wave (b) Reflected wave.

Figure 8.16 illustrates the analysis for an elliptical boundary. When the frontal

wave reaches M in Figure 8.16a, it has the shape of a circular arc with center at the initial focus F_1. Let r denote its radius. A tiny portion of the wave around M is reflected from the elliptical wall according to the reflection property common to all waves and particles: the angle of reflection is equal to the angle of incidence. Because the boundary is an ellipse, the reflected portion will move directly toward the second focus along the dashed line in Figure 8.16a, and the reflected wave will have the shape of another curve passing through M (Figure 8.16b). To determine the shape, consider another tiny portion of the frontal wave that has been reflected earlier, say from point N. That portion is also directed toward the second focus and is located at some position P on the reflected wave. We will show that the distance $R = PF_2$ from P to F_2 is independent of P. The two portions move with the same speed, so $F_1N + NP = F_1M = r$. But $F_1N + NF_2 = 2a$, the length of the major axis of the ellipse. Subtracting the last two equations we find $NF_2 - NP = 2a - r$. But $NF_2 - NP = PF_2 = R = 2a - r$, a constant independent of P. Thus the frontal wave and the reflected wave are parts of two focal circles of respective radii r and R. But $r + R = 2a$, so these circles are circular directrices for the ellipse.

If the pebble is dropped at one focus of a hyperbolic pool, the reflected wave is again a circular arc expanding away from the second focus. And if the pebble is dropped at the focus of a parabolic pool, the reflected wave will be a floating directrix moving away from the vertex of the parabola. For animation, see www.mamikon.com/Pool/FocalWave.html.

8.7 EXTENDED ECCENTRICITY PROPERTIES OF CONICS

All three types of conics are often described by a classical eccentricity property (Figure 8.17a), as the locus of a moving point P whose distance d_F from a fixed point F (a focus) is a constant e times its distance d_D from a fixed line D (the directrix). The constant ratio e is the eccentricity.

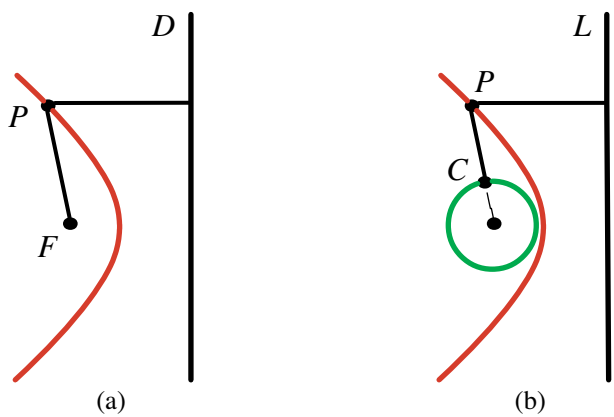


Figure 8.17: (a) Classical eccentricity property of all conic sections: $d_F = ed_D$ for each P. (b) Extended eccentricity property: $d_C = ed_L$ for each P.

The classical eccentricity property is written as

$$d_F = ed_D, \tag{8.11}$$

and the conic is an ellipse if $0 < e < 1$, a parabola if $e = 1$, and a hyperbola if $e > 1$. Now we replace the focus F by a circle C and we replace the directrix D by an arbitrary line L coplanar with C, and extend the eccentricity property (8.11) as follows.

Theorem 8.5. *Given a circle C, a coplanar line L, and a constant $e > 0$. For any point P in the plane of L and C, let d_L and d_C denote the shortest distances from P to L and C, respectively. Then the locus of all P such that*

$$d_C = ed_L \tag{8.12}$$

is a conic with eccentricity e and a focus at the center of C. The same is true if d_C is replaced by D_C, the longer distance from P to C.

Proof. Denote the center of C by F and its radius by r. If P is outside C there are two distances d_C, given by $d_F \pm r$. The shorter is $d_C = d_F - r$, and (8.12) is equivalent to $d_F - r = ed_L$, or $d_F = ed_L + r = e(d_L + r/e)$. But $d_L + r/e = d_D$, the distance from P to a line D parallel to L and at distance r/e from L. Thus, (8.12) is equivalent to $d_F = ed_D$, and by (8.11) the locus is a conic with focus F, eccentricity e, and directrix D.

If D_C is the longer distance to C, then $D_C = d_F + r$, and the same argument gives $d_F = e(d_L - r/e) = ed'_D$, where d'_D is the distance from P to the other line D' parallel to L and at distance r/e from L. In other words, if D_C denotes the longer distance to C the locus is again a conic with focus F and eccentricity e, but with directrix D'. Similarly, for points P inside C the same conclusion (8.12) is valid.

When Theorem 8.5 is specialized so circle C shrinks to its center F, then (8.12) becomes $d_F = ed_L$, which is the classical result (8.11) with $L = D$. Theorem 8.5 is the special case $c = 0$ of the following more general result, whose proof is omitted.

Theorem 8.6. *Given a circle C, a coplanar line L, a constant $e > 0$ and a constant $c \geq 0$, for any point P in the plane of L and C let d_L and d_C denote the shortest distances from P to L and C, respectively. Then the locus of all P such that*

$$|d_C \pm ed_L| = c \tag{8.13}$$

is a conic with eccentricity e, and a focus at the center of C. The same is true if d_C is replaced by D_C, the longer distance from P to C.

When C is replaced by its center F we obtain the following

Corollary 8.1. *The locus of all P such that*

$$|d_F \pm ed_L| = c \tag{8.14}$$

is a conic with eccentricity e, a focus at F, and directrix at distance c/e from L.

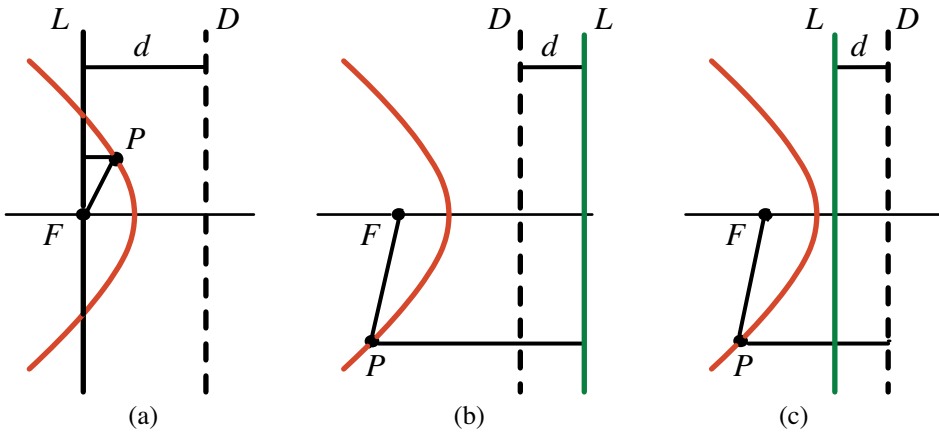


Figure 8.18: Examples of conics with (a) $d_F + ed_L = ed$, (b) $ed_L - d_F = ed$, and (c) $d_F - ed_L = ed$.

Corollary 8.1 is illustrated by the examples in Figure 8.18.

NOTES ON CHAPTER 8

Except for Section 8.6 on reflected waves, the material in this chapter first appeared in [25]. The material in Section 8.6 has not been previously published.

In this chapter we learned that the simple idea of replacing each focus of a conic by a focal circle has profound consequences. It allows us to obtain new characteristic properties of central conics and to extend them to a parabola.

The classical characterization of an ellipse as the locus of points whose sum of focal distances $f_1 + f_2$ is constant, and the hyperbola as the locus of points whose absolute distance $|f_1 - f_2|$ is constant, has been generalized to a common bifocal property $|d_1 \pm d_2|$ is constant, where d_1 and d_2 are the shortest distances from a point to the focal circles. By allowing the radius of one of the focal circles to become infinite, we obtained corresponding properties for the parabola.

We also introduced special pairs of focal circles, called circular directrices, which provide equidistant properties for central conics analogous to the classical focus-directrix equidistant property for the parabola. A circular directrix also provides an auxiliary tool to treat a variety of problems in a unified and elementary way.

The classical locus description of all three types of conics in terms of eccentricity has also been generalized by using focal circles and floating focal lines.

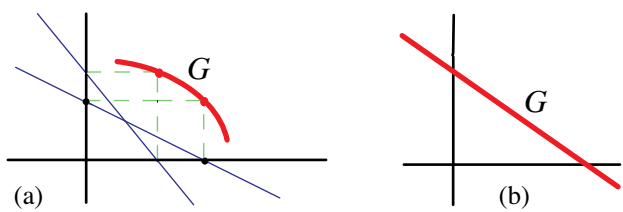
Chapter 9

TRAMMELS

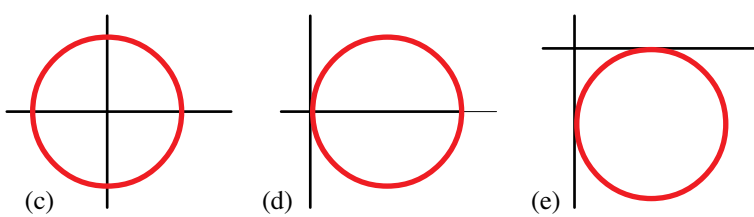
These problems can be easily solved by the methods developed in this chapter. The reader may wish to try solving them before reading the chapter.

Figure (a) shows a curve G in the xy plane. Drop a perpendicular from each point on G to each coordinate axis, and pass a line through the feet of the two perpendiculars, as shown by the two examples in (a). When this geometric construction is done on graph paper, it leads to the following educational activity:

Fold the paper along each of the lines joining the axes, so each line appears as a crease. By choosing many points on G and repeating this process, the positions of the crease gradually reveal a curve having these creases as tangent lines. We call this curve the *envelope* generated by G.



If G is the line shown in (b)*, show that the envelope is a parabola.*



Find a cartesian equation for the envelope when G is a circle: centered at the origin as in (c)*, tangent to the y axis with center on the x axis as in* (d)*, tangent to both coordinate axes as in* (e)*.*

CONTENTS

9

The chapter begins with an elementary treatment of a standard trammel (trammel of Archimedes), a line segment of fixed length whose ends slide along two perpendicular axes. During the motion, points on the trammel trace ellipses, and the trammel produces an astroid as an envelope that is also the envelope of the family of traced ellipses. Two generalizations are introduced: a zigzag trammel, obtained by dividing a standard trammel into several hinged pieces, and a flexible trammel whose length may vary during the motion. All properties regarding traces and envelopes of a standard trammel are extended to these more general trammels. Applications of zigzag trammels are given to problems involving folding doors. Flexible trammels provide not only a deeper understanding of the standard trammel but also a new simple solution of a classical problem of determining the envelope of a family of straight lines. They also reveal unexpected connections between various classical curves; for example, the cycloid and the quadratrix of Hippias, known from antiquity.

9.1 INTRODUCTION. STANDARD TRAMMEL

Figure 9.1a shows a line segment of fixed length whose ends slide along two perpendicular axes. It can be realized physically as a sliding ladder or as a sliding door moving with its ends on two perpendicular tracks. During the motion, a given point on the segment traces an ellipse with one quarter of the ellipse in each quadrant (Figure 9.1b), so this device is called an *ellipsograph*, a mechanism for drawing an ellipse.

This particular ellipsograph was known to ancient Greek geometers and is often called the trammel of Archimedes [70, p. 3], but we are not aware of any historical evidence that suggests who invented it. We refer to it as a *standard trammel* to distinguish it from two generalizations introduced in this chapter: a *zigzag trammel*, obtained by dividing a standard trammel into several hinged pieces (Section 9.8), and a *flexible trammel* whose length may vary (Section 9.9).

The ellipsograph feature alone makes the trammel an important object of study, but it has other interesting properties discovered with methods of calculus that

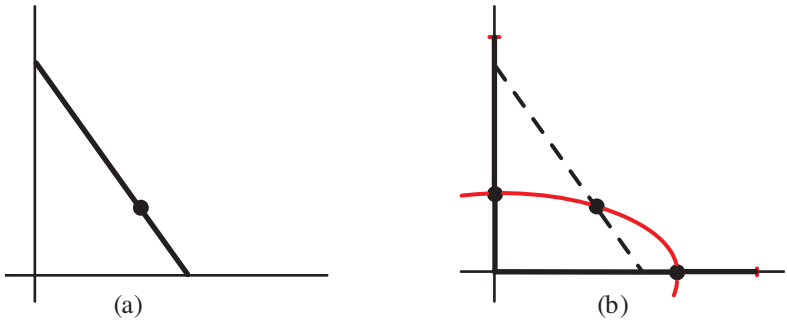


Figure 9.1: (a) Standard trammel. Its endpoints move along perpendicular axes. (b) Each point on the trammel traces an ellipse.

deserve to be better known. This chapter describes known properties of the standard trammel as well as new ones that can be studied by simple geometric methods that do not require calculus. The generalization to flexible trammels increases our understanding of these properties and also leads to engaging classroom activities.

9.2 ELLIPSE TRACED BY A POINT ON A STANDARD TRAMMEL

To see why a point on a trammel AB of fixed length L traces an ellipse, we show first that the midpoint M of AB traces a circle, as suggested by Figure 9.2a. Figure 9.2b shows the segment OM, together with perpendiculars from M to each axis, dividing the large right triangle AOB into four congruent smaller right triangles, each having hypotenuse of common length $OM = AM = MB = L/2$. Therefore, M always lies on a circle with center at O and radius $L/2$. This also follows by completing the rectangle $OACB$ in Figure 9.2c.

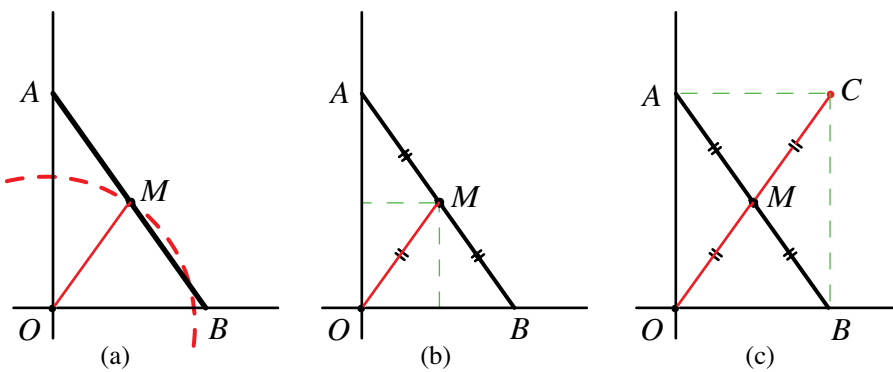


Figure 9.2: (a) The midpoint M traces a circle of radius OM. (b) The triangle AOB divided into four congruent right triangles. (c) The diagonals of a rectangle are equal and bisect each other.

The circle of radius $L/2$ has cartesian equation

$$(\frac{x}{L/2})^2 + (\frac{y}{L/2})^2 = 1. \tag{9.1}$$

This is a special case of the general equation of an ellipse,

$$(\frac{x}{a})^2 + (\frac{y}{b})^2 = 1, \tag{9.2}$$

traced by a point $E = (x, y)$ on the trammel that divides it into segments of lengths $a = AE$ and $b = EB$, as indicated in Figure 9.3a. We deduce (9.2) from (9.1) by relating the coordinates (x, y) of E with the coordinates (X, Y) of the midpoint M that satisfy (9.1). Similar triangles in Figure 9.3a reveal that $x/a = X/(L/2)$, and $y/b = Y/(L/2)$, so (9.2) follows from (9.1). The same analysis holds if E is anywhere on the line through the trammel, even if E is *outside* the trammel as in Figure 9.3b. In all cases:

The semiaxes of the ellipse are the distances from E to the endpoints A and B.

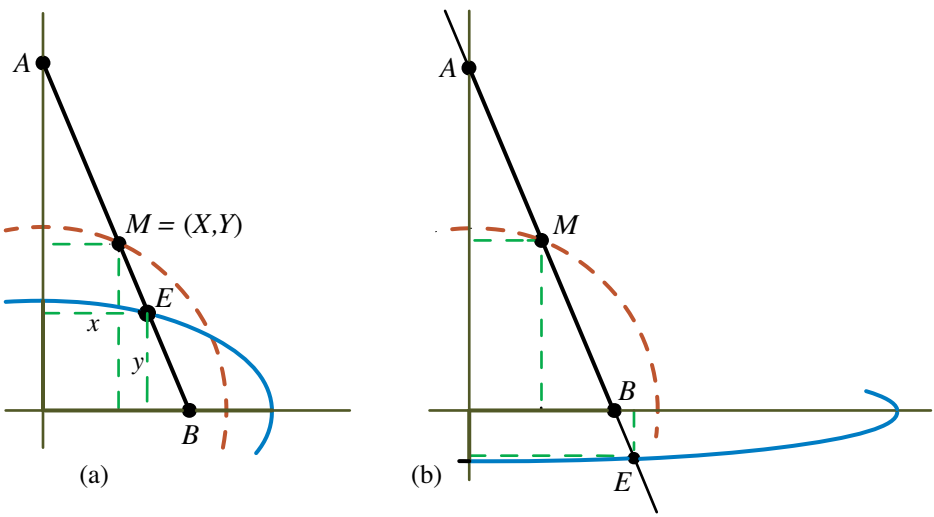


Figure 9.3: The coordinates (x, y) of E are related to those (X, Y) of M as follows: $x/a = X/(L/2)$, and $y/b = Y/(L/2)$, where $a = AE$ and $b = EB = |L - a|$. In (a), E lies between A and B; in (b), E lies outside AB but on the line through AB.

9.3 ASTROID AS THE ENVELOPE OF A STANDARD TRAMMEL

Next we show geometrically that a trammel of length L is always tangent to an *astroid*, a hypocycloid traced by a point P on the circumference of a circle that rolls without slipping inside a larger circle of radius four times as great. (See Chapter 2.) Figure 9.4a shows the larger circle of radius L centered at the origin, and the smaller circle of radius $L/4$ rolling inside it.

Let C denote the point of tangency of the two circles. The midpoint M of OC is on the smaller circle as shown. Triangle MPC is inscribed in a semicircle with diameter MC, so the angle MPC is a right angle, as indicated. The line AB drawn through PM, perpendicular to CP, is also tangent to the astroid at P because C is the center of instantaneous rotation of the smaller circle as it rolls inside the larger circle. Therefore, if we show that the length of AB does not change as P moves along the astroid, this will show that AB is a trammel. To do this it suffices to show that M is also the midpoint of AB, and that the triangle OMA is isosceles.

Refer to Figure 9.4b. The two circular arcs CL and CP have equal lengths because the smaller circle rolls along the larger. The arc CL has length $L\gamma$, where γ is the central angle of the larger circle subtending arc CL. Similarly, CP has length $(L/4)\beta$, where β is the central angle of the smaller circle subtending CP. The arcs have equal length so $\beta = 4\gamma$, as indicated in Figure 9.4b. But the inscribed angle CMP on the smaller circle is half the central angle, or 2γ. This is also the vertical angle OMB. A vertical line through M makes angle γ with OC, so it bisects

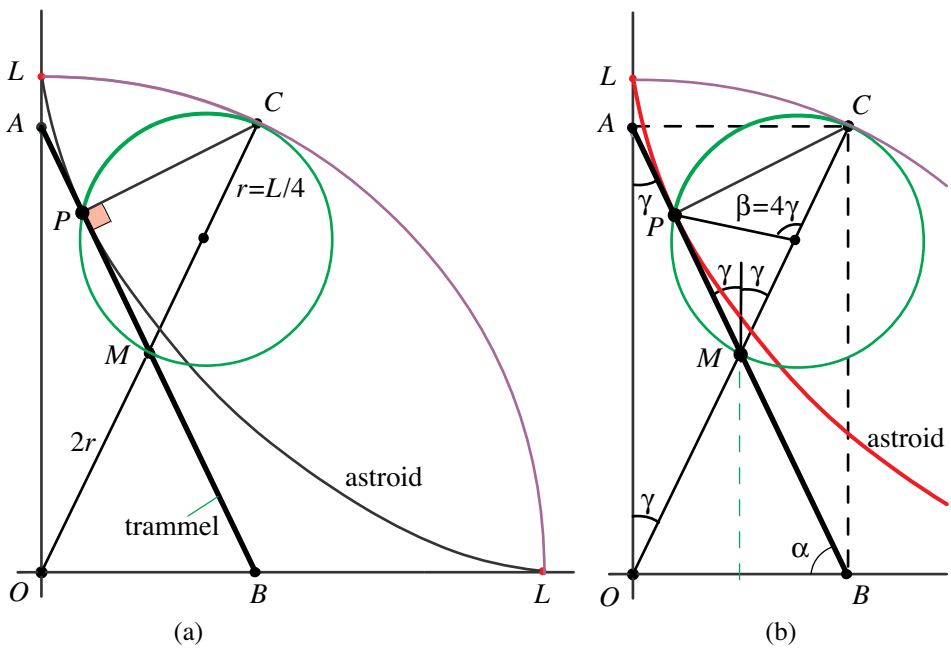


Figure 9.4: (a) A point P on the circle of radius $L/4$ traces an astroid as it rolls inside a circle of radius L. (b) Proof that AB has constant length equal to L.

CMA as shown. Hence the triangle OMA is isosceles, with each base angle equal to γ. This implies that OMB is also isosceles with base angles α complementary to γ. Therefore $AM = MB = MO = L/2$, which shows that AB has fixed length L, so it is a standard trammel, and we have already seen that it is always tangent to the astroid.

This property is illustrated in Figure 9.7a, which shows various positions of the

trammel as it slides along the coordinate axes in all four quadrants. It is described by saying that:

The envelope of a moving trammel of fixed length is an astroid.

9.4 EQUATIONS FOR ELLIPSE, TRAMMEL, AND ASTROID

Figure 9.5 shows an ellipse with semiaxes a and b and two concentric circles with radii a and b. A line from the origin making an angle α with the x axis is related to the coordinates of a point $E = (x, y)$ on the ellipse by the parametric equations

$$x = a\cos\alpha, \quad y = b\sin\alpha. \tag{9.3}$$

The angle α in (9.3) is called the *eccentric angle of the ellipse* (see [1, p. 522]).

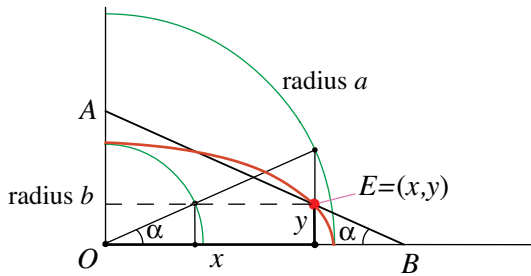


Figure 9.5: Angle of inclination of a trammel is also the eccentric angle of traced ellipse.

Figure 9.5 also shows a line through E making the same angle α with the x axis. It is easy to show that AB has length $a + b$. We simply note that $AE\cos\alpha = x$ so according to (9.3), $AE = a$. Similarly, $BE\sin\alpha = y$, giving $BE = b$. Therefore segment AB is a trammel whose angle of inclination is the eccentric angle.

In Figure 9.6a, a line AB of length L makes an angle α with the x axis and is tangent to the astroid at a point P with coordinates (x, y). First we note that every point on the trammel satisfies the linear cartesian equation

$$\frac{x}{L\cos\alpha} + \frac{y}{L\sin\alpha} = 1 \tag{9.4}$$

because $L\cos\alpha$ and $L\sin\alpha$ are the x and y intercepts of the trammel.

To find a cartesian equation for the astroid we first determine parametric equations expressing the coordinates (x, y) of each point on the astroid in terms of α and L. From Figure 9.6a, we see that

$$x = AP\cos\alpha, \quad y = BP\sin\alpha. \tag{9.5}$$

To express AP and BP in terms of L and α, we recall from Figure 9.4b that point C of the rolling circle is a vertex of rectangle $OACB$. From the right triangle AOB

in Figure 9.6a we see that $AC = OB = L\cos\alpha$, and from the right triangle APC we have $AP = AC\cos\alpha = L\cos^2\alpha$. Similarly we find $PB = L\sin^2\alpha$. Using these in (9.5) we see that each point (x, y) on the astroid satisfies the parametric equations

$$x = L\cos^3\alpha, \quad y = L\sin^3\alpha. \tag{9.6}$$

Using (9.6) together with $\cos^2\alpha + \sin^2\alpha = 1$ we obtain a cartesian equation of the astroid:

$$(\frac{x}{L})^{2/3} + (\frac{y}{L})^{2/3} = 1. \tag{9.7}$$

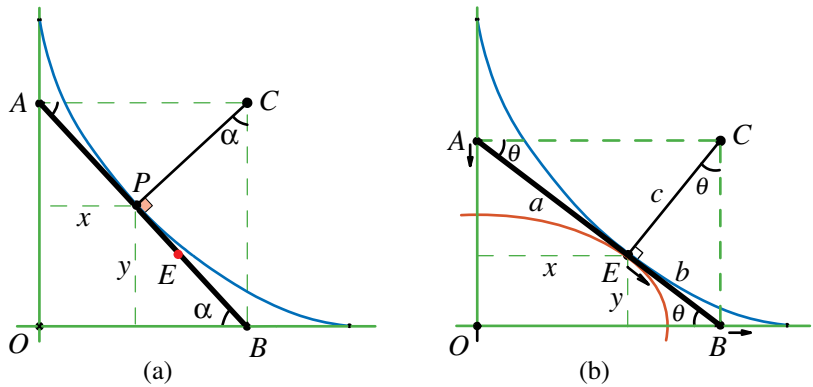


Figure 9.6: (a) Finding equations for an astroid. (b) The trammel is simultaneously tangent to the ellipse and to the astroid at E when inclined at a special angle θ given by (9.8).

9.5 COMMON TANGENCY OF TRAMMEL, ELLIPSE, AND ASTROID

In Figure 9.6a, the point P with coordinates (9.5) is on the trammel (9.4) and on the astroid (9.7). The trammel is always tangent to the astroid but in general it intersects the ellipse described by (9.3) at two points as in Figure 9.5. As the trammel moves, the two points will coincide for some critical angle α, say $\alpha = \theta$ as shown in Figure 9.6b, and in this position the trammel is simultaneously tangent to the ellipse and to the astroid at a point E with coordinates $(a\cos\theta, b\sin\theta)$. The following theorem determines θ in terms of the semiaxes a and b of the ellipse.

Theorem 9.1. *The trammel is simultaneously tangent to the ellipse and to the astroid when its angle of inclination θ with the x axis satisfies*

$$\tan\theta = \sqrt{\frac{b}{a}}. \tag{9.8}$$

Proof. In Figure 9.6b, $c = CE$, and triangles AEC and BEC are similar, so $a/c = c/b$, or $c^2 = ab$. The triangle BEC reveals that $\tan\theta = b/c = b/\sqrt{ab} = \sqrt{b/a}$, which proves (9.8).

Note that C in Figure 9.6b is also the center of instantaneous rotation of the trammel AB. As A moves downward, B moves to the right, and E moves in the direction of AB as indicated. This also shows that the trammel is simultaneously tangent to the ellipse and to the astroid at E. As a and b vary, keeping $a+b$ constant, the corresponding ellipses are tangent to the astroid. This property, illustrated in Figure 9.7b, is described as:

The envelope of a family of ellipses with $a + b$ constant is an astroid.

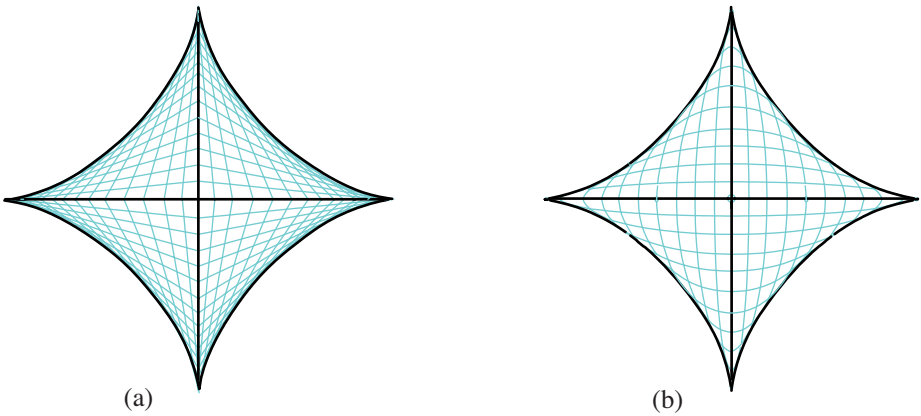


Figure 9.7: Astroid as the common envelope of (a) a moving trammel, and (b) a family of traced ellipses.

A generalization of Theorem 9.1 to flexible trammels is given in Section 9.10, leading to a common envelope for a flexible trammel and the curves traced by its points.

9.6 AREA OF AN ELLIPTICAL SECTOR

The trammel is shown in Figure 9.8 as a segment AB of length $a + b$ in the first quadrant inclined at an angle α with the x axis. The ellipse is traced by E, where AE has length a and EB has length b. As noted earlier, α is the eccentric angle of the ellipse. Let $S(\alpha)$ denote the area of the shaded elliptical sector OE_0E. Here, E_0 denotes the position of E when the trammel is horizontal. Then we have the following simple result that we shall deduce without calculus:

Theorem 9.2. *The elliptical sector OE_0E has area*

$$S(\alpha) = \frac{1}{2}ab\alpha. \tag{9.9}$$

Proof. The elliptical sector can be obtained from a circular sector with central angle β and radius b (Figure 9.8b) by horizontal dilation by the factor a/b (Figure 9.8c). The circular sector has area $b^2\beta/2$, so the dilated sector has area a/b times as much or $ab\beta/2$. But Figure 9.8d shows that $\beta = \alpha$ because both are base angles of

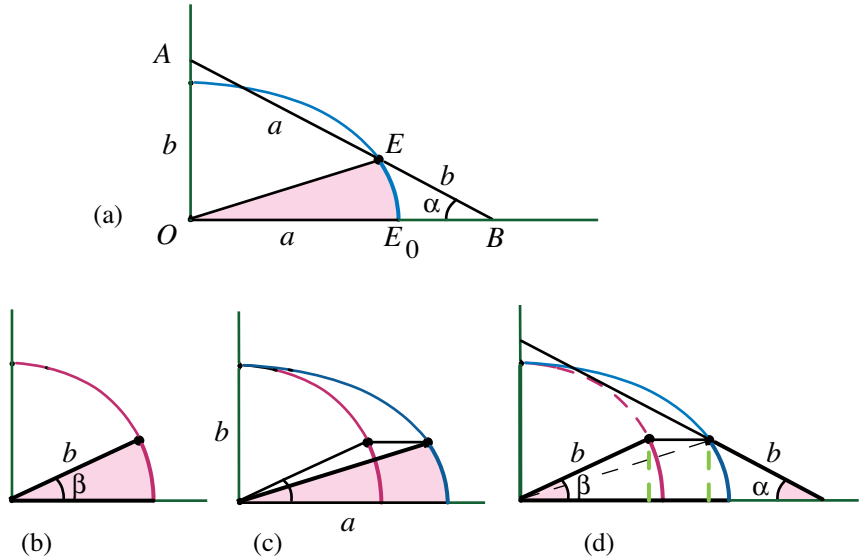


Figure 9.8: (a) Elliptical sector OE_0E determined by trammel inclined at angle α. (b)-(d) Proof that the area of the sector is $ab\alpha/2$.

congruent right triangles having hypotenuses of equal length b and equal altitudes. Therefore, this simple geometric argument gives (9.9).

When $a = b$ the ellipse is a circle of radius b and (9.9) gives the area of a circular sector in terms of the eccentric angle, which now equals the central angle.

The right member of (9.9) is linear in α, so the area of the sector of the ellipse between two values of α, say $0 < \alpha_1 < \alpha_2 \leq 2\pi$, is $ab(\alpha_2 - \alpha_1)/2$. Thus the area of a more general elliptical sector, such as that shown in Figure 9.9a is given by:

$$\textit{Area of general elliptical sector} = \frac{1}{2}ab\varphi,$$

where $\varphi = \alpha_2 - \alpha_1$ denotes the angle between the two trammel positions, which is also the change in eccentric angle. In particular, if φ is a right angle (two perpendicular trammel positions) the area of the sector is $\pi ab/4$, one quarter of the area of a full ellipse. An example is shown in Figure 9.9b.

9.7 AREA OF AN ASTROIDAL SECTOR

It is known (see Chapter 2) that the area of the region enclosed by the astroid in Figure 9.10a is 6 times the area of the smaller circular disk that traces the astroid, so in each quadrant the astroidal area is 3/2 times that of the rolling disk.

Figure 9.10b shows the region in the first quadrant composed of a quarter of a disk of radius $2r$ and area πr^2, together with two congruent shaded regions between the disk and the astroid. The rolling disk has radius r. We will show that each

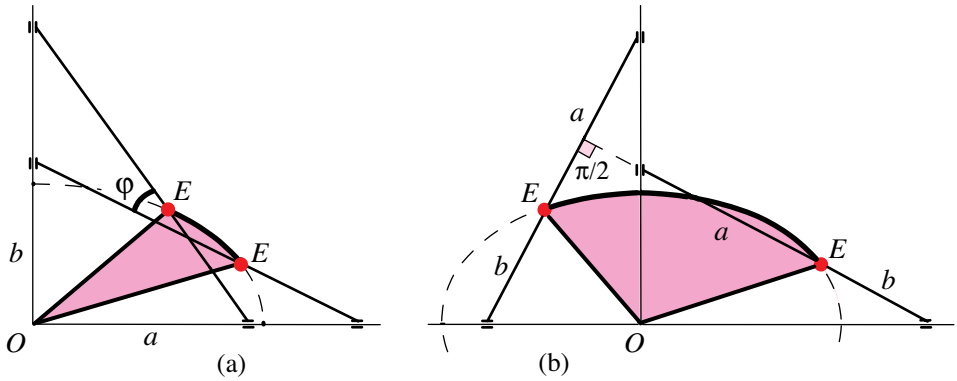


Figure 9.9: (a) Area of a general elliptical sector. (b) Special case with $\varphi = \pi/2$.

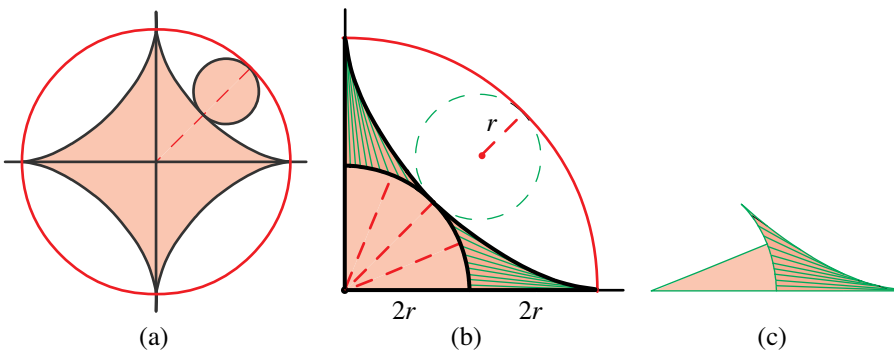


Figure 9.10: (a) The region enclosed by an astroid has area 6 times that of the rolling circle. (b)The region in the first quadrant below the astroid has area $3\pi r^2/2$. (c) The area of an astroidal sector is equal to that of a circular sector.

shaded region has area $\pi r^2/4$, so the two regions together with the quarter disk have area $3\pi r^2/2$. The surprising equality of the area of an astroidal sector with that of a circular sector is illustrated in Figure 9.10c. Before verifying the formulas, we mention an application to sliding doors.

Comparing a hinged door with a sliding door.

From Figure 9.10b we see that a door of length $4r$ hinged at the origin sweeps out a floor space of area $4\pi r^2$. By comparison, a door of the same length sliding along perpendicular tracks sweeps out a floor space of area $3\pi r^2/2$, which is 3/8 of $4\pi r^2$, an impressive saving of 62.5%. Examples involving sliding doors that fold are discussed in Section 9.8.

The area of an astroidal sector will be deduced from a more general result, illustrated in Figure 9.11a, where the shaded region is swept by the portion cut off by the quarter disk of radius $2r$ as the trammel moves from a horizontal position

to a general angle of inclination α as shown. We call this region a *general astroidal sector* and denote its area by $A_r(\alpha)$. This region is an example of a *tangent sweep*,

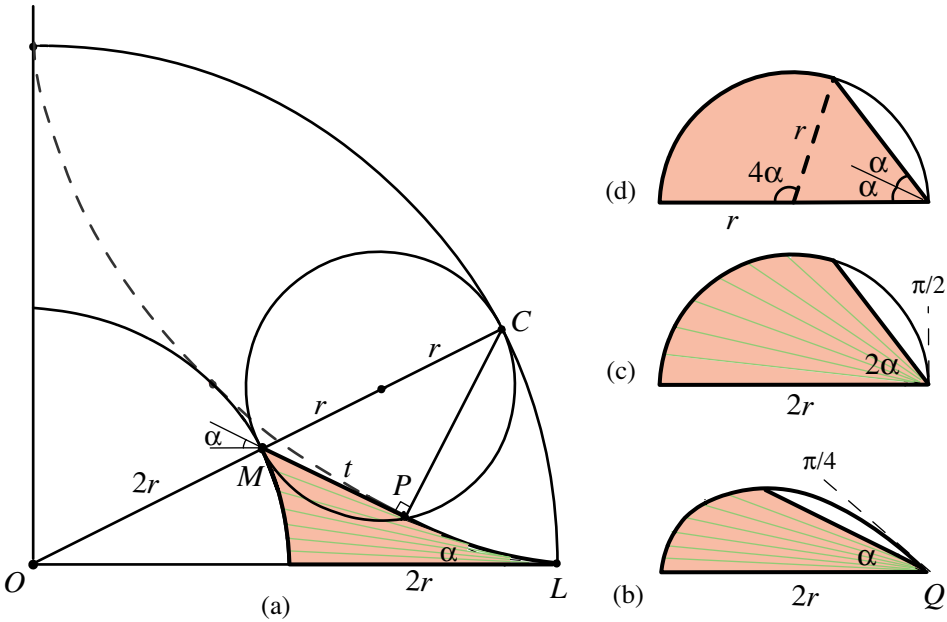


Figure 9.11: Calculating the area of a general astroidal sector.

and Figure 9.11b shows the corresponding *tangent cluster*, in which the tangent trammel segments have been translated so that each point of tangency is moved to a common point Q. Mamikon's sweeping tangent theorem (Chapter 1) tells us that the tangent sweep and the tangent cluster have equal areas. Using Figure 9.11 we shall prove:

Theorem 9.3. *The area of the general astroidal sector is given by*

$$A_r(\alpha) = r^2\alpha + \frac{1}{4}r^2 \sin 4\alpha. \tag{9.10}$$

Proof. Figure 9.11a shows the larger circle of radius L centered at the origin, and the smaller circle of radius $r = L/4$ rolling inside it, with P denoting the point on the astroid when the trammel has turned through angle α from the horizontal. Point M, the midpoint of OC, is also on the smaller circle as shown, and t denotes the length of the tangent segment PM. The argument used in Figure 9.4a shows that the angle PMC in Figure 9.11a is equal to 2α, so $t = 2r\cos 2\alpha$. Therefore the tangent cluster shown in Figure 9.11b is bounded by a portion of a rosette whose polar coordinates (t, α) relative to the origin at Q satisfy $t = 2r\cos 2\alpha$.

The area of the tangent cluster in (b) can be easily calculated using calculus. It can also be determined without calculus by noting that the area is half that of the

portion of a semicircular disk enclosed by the inscribed angle 2α in (c). This area, in turn, is equal to that of a circular sector of radius r subtended by a central angle 4α, which is $2r^2\alpha$, plus that of an isosceles triangle with legs r and vertex angle 2α, which is $\frac{1}{2}r^2 \sin 4\alpha$, so the area of the region in (b), and hence that of the general astroidal sector in (a), is given by (9.10).

When $\alpha = \pi/4$ this gives $\pi r^2/4$ for each astroidal sector in Figure 9.10b, and proves the surprising area equality in Figure 9.10c.

Application: Area of the region swept by a portion of a trammel.

Figure 9.12a shows a trammel of length $a + b$ simultaneously tangent to an ellipse with semiaxes a and b and to an astroid at point E where it makes a critical angle θ with the x axis as determined by (9.8). As applications of (9.9) and (9.10) we shall calculate, in terms of a, b, and θ, the area $A(\theta)$ of the region swept out by the trammel segment of length b as the trammel slides from a horizontal to a vertical position. The region in question, under the ellipse and the astroid, is shown shaded in Figure 9.12a. We build this region in four steps as shown in Figures 9.12b-e.

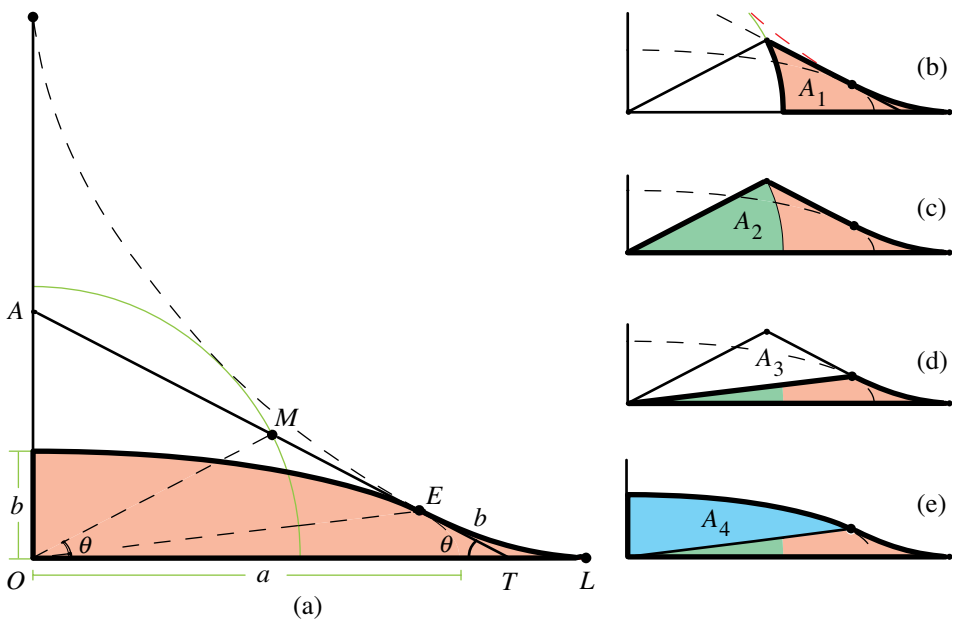


Figure 9.12: Region (a) swept by a portion of a trammel of length b constructed in (b)-(e).

The shaded region in Figure 9.12b, whose area we denote by A_1, is like the astroidal sector in Figure 9.10a cut off by the circle of radius $2r = (a+b)/2$. Figure 9.12c shows the astroidal sector of Figure 9.12b adjacent to a circular sector whose area we denote by A_2. In Figure 9.12d a triangle of area A_3 has been removed from the region of Figure 9.12c. In Figure 9.12e an elliptical sector of area A_4 is added to the region of Figure 9.12d, thus completing the shaded region in Figure 9.12a.

Hence

$$A(\theta) = A_1 + A_2 - A_3 + A_4, \tag{9.11}$$

and now we calculate each area A_i separately.

From (9.10) we find $A_1 = A_r(\theta) = r^2\theta + \frac{1}{4}r^2 \sin 4\theta$. Point M is the midpoint of the trammel so OM also makes an angle θ with the x axis, hence the circular sector in Figure 9.12c has area $A_2 = 2r^2\theta$. In Figure 9.12d the area of the triangle is the difference of areas of triangles OMT and OET in Figure 9.12a. They have a common base $(a+b)\cos\theta$ and respective altitudes $2r\sin\theta$ and $b\sin\theta$. Hence we see that

$$A_3 = (a+b)(r - \frac{b}{2})\sin\theta\cos\theta = r(2r-b)\sin\,2\theta.$$

Finally, from (9.9) we infer that the elliptical sector in Figure 9.12e has area $A_4 = \pi ab/4 - ab\theta/2$. Using these values in (9.11) we find $A(\theta)$ as the following Corollary of Theorems 9.1, 9.2 and 9.3:

Corollary.

$$A(\theta) = 3r^2\theta + \frac{r^2}{4}\sin 4\theta - r(2r-b)\sin\,2\theta + \frac{ab}{2}(\frac{\pi}{2} - \theta), \tag{9.12}$$

where $r = (a+b)/4$, *and* θ *is given by* (9.8).

9.8 ZIGZAG TRAMMEL

Start with a standard trammel ZT of length L regarded as a rod initially on the positive x axis, with Z at $(0,0)$ and T at $(L,0)$, as shown in Figure 9.13a. Divide

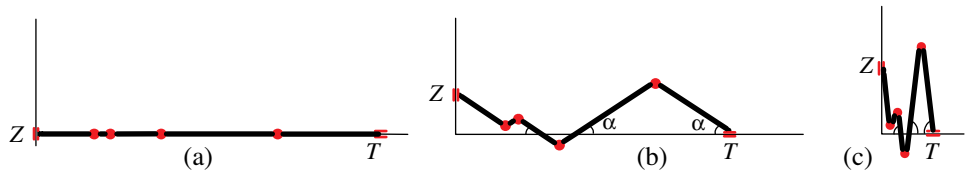


Figure 9.13: (a) Standard trammel of length L divided into rods. (b) and (c). Zigzag trammel of length L, with all rods making equal angles with the x axis.

ZT into a finite number of hinged rods of arbitrary lengths whose sum is L. Now let the free ends T and Z slide along the respective coordinate axes as indicated in Figures 9.13b, c, with all rods making equal angles with the x axis. We call this a *zigzag trammel* if the number of rods is greater than 1. In the examples shown, the angle α is allowed to increase from 0 to $\pi/2$ so the rightmost rod will stay in the first quadrant.

For $\alpha > 0$, an ant starting at T and walking along the zigzag trammel toward Z decreases its x coordinate monotonically, but its y coordinate is piecewise monotonic, alternately increasing or decreasing as it crosses the hinges.

A line parallel to the rightmost rod RT is called a *zig* line, while those parallel to the rod adjacent to RT are called *zag* lines. Because all rods make equal angles with the x axis, they generate a finite set of parallel zig lines and another finite set of parallel zag lines. Each hinge is the intersection of a zig line and a zag line, as indicated in Figure 9.14a. The configuration is like a hinged latticework in which two adjacent rods determine a parallelogram.

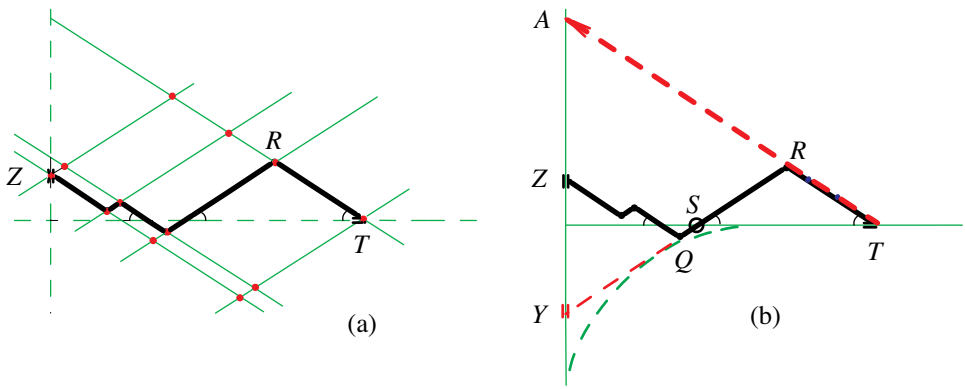


Figure 9.14: (a) Parallel zig lines and parallel zag lines through the hinges of a zigzag trammel. (b) A line through the rightmost rod RT intersects the y axis at A.

In Figure 9.14b, A denotes the point at which a line through the rightmost rod RT intersects the y axis. By the parallel construction suggested by Figure 9.14a, it follows that the length of AT is the sum of the lengths of the rods of the zigzag trammel. Consequently, we have the following property:

Property 1. *As a zigzag trammel ZT of length L slides along the axes, a line segment AT through the rightmost rod RT, from the y axis to T, is a standard trammel of length L sliding along the same axes, regardless of the number of rods and their relative sizes.*

Applications to folding doors.

Zigzag trammels can be realized physically as folding doors. Examples are those treated in [63], where the left endpoint Z is kept fixed. The seemingly unexpected appearances of the same astroid and various ellipses associated with bifold closet doors treated in [63] are easily explained as consequences of Property 1. In [63] the floor area swept out by the right panel of a bifold closet door was found using calculus. That result follows directly from the discussion related to Figures 9.10b and 9.10c. It is also a special case of (9.12) when $b = a = L/2$ and $\theta = \pi/4$. As we observed earlier, using a one-panel door sliding along two perpendicular tracks instead of a hinged door results in a saving of 62.5% in the swept floor area. Using a bifold closet door as is done in [63] with Z fixed saves only an additional 6.25%. We can do much better if we allow Z to move upward along the vertical axis while

keeping the hinge on the horizontal track. Then the left panel is a trammel of half the length that sweeps out one-quarter as much floor space. When the hinge reaches the vertical axis and then moves upward, the right panel covers the same floor space already covered by the left panel. The total saving in floor space is now 29/32 of that swept by a non-folding hinged door, or a whopping 90.63%.

If the two rods joining alternate hinges of a zigzag trammel always have equal lengths, they form an isosceles triangle whose vertex is at the intermediate hinge. This configuration occurs when closet doors are divided into two or more hinged panels that fold or unfold as the door is opened or closed. Equation (9.12) can be used in particular to calculate the floor area swept by the rightmost panel of an n-panel door of equal sizes by taking $b = L/n$ and $a = L(n-1)/n$. We omit the details.

A zigzag trammel has other interesting properties. For example, consider the rod RQ in Figure 9.14b adjacent to the rightmost rod RT. Let S denote the point where RQ or its extension to Y on the y axis intersects the x axis. Then we have:

Property 2. *As a zigzag trammel ZT of length L slides along the axes, the point S remains on the x axis and SY has constant length $L - 2RT$, so SY is a standard trammel sliding along the same axes.*

Proof. During the entire motion, TRS is an isosceles triangle with $RS = RT$, so S remains on the x axis. Also $SY = YR - SR = (L - RT) - SR = L - 2RT$.

The same analysis applies to each of the rods of a zigzag trammel. A line through the rod extended to meet the y axis produces a standard mini-trammel whose ends slide along the same axes. Consequently, each point on such a mini-trammel traces its own ellipse. It also has its own astroid as an envelope. One of them is shown by the dashed curve in Figure 9.14b. The next theorem tells how to determine the length of each mini-trammel.

Theorem 9.4. *Consider a zigzag trammel of length L with rods of lengths $z_1, z_2, ..., z_n$, labeled from right to left. Then the segment between the x and y axes of the line through the ith rod is a standard trammel whose length L_i is given as follows:*
for odd i, (zig lines)

$$L_1 = L, \;\; L_{2k+1} = L - 2(z_2 + z_4 + \cdots + z_{2k}), \;\; k \geq 1,$$

and for even i, (zag lines)

$$L_{2k} = L - 2(z_1 + z_3 + \cdots + z_{2k-1}), \;\; k \geq 1.$$

An inductive proof can be given, which is omitted. The formulas for L_1 and L_2 to begin the induction are contained in Properties 1 and 2, respectively.

A knowledge of the length of a mini-trammel determines its corresponding astroidal envelope, and also determines the semiaxes of the ellipse traced by a point on the line through each rod. The semiaxes are the distances from the tracing point to the endpoints of the corresponding mini-trammel. In particular the semiaxes a_k

and b_k of the ellipse traced by the hinge joining rod k with rod $k+1$ are given by

$$a_k = L - (z_1 + \cdots + z_k), \;\; b_k = |L_k - a_k|,$$

where L_k is the trammel length determined by Theorem 9.4.

For animations showing zigzag trammels in motion see the web site

www.mamikon.com/Trammel.html.

9.9 FLEXIBLE TRAMMEL

An astroid is the common envelope of a moving standard trammel and a family of traced ellipses (Figure 9.7). Now we introduce trammels of variable length that provide a new solution of a classical problem of finding the envelope of a family of straight lines. In Theorem 9.7 we will show that this is also the envelope of a family of curves traced by special points on these lines. Sections 9.11 and 9.12 provide many interesting examples.

Figure 9.15a shows the first quadrant segment of a line with cartesian equation

$$\frac{x}{m} + \frac{y}{n} = 1. \tag{9.13}$$

Here m and n are the x and y intercepts of the line. When both intercepts are nonzero they can be regarded as legs of a right triangle of lengths $|m|$ and $|n|$ having hypotenuse of length $\sqrt{m^2+n^2}$. If m^2+n^2 is constant, the line segment joining the axes will have constant length and we obtain a standard trammel of that length.

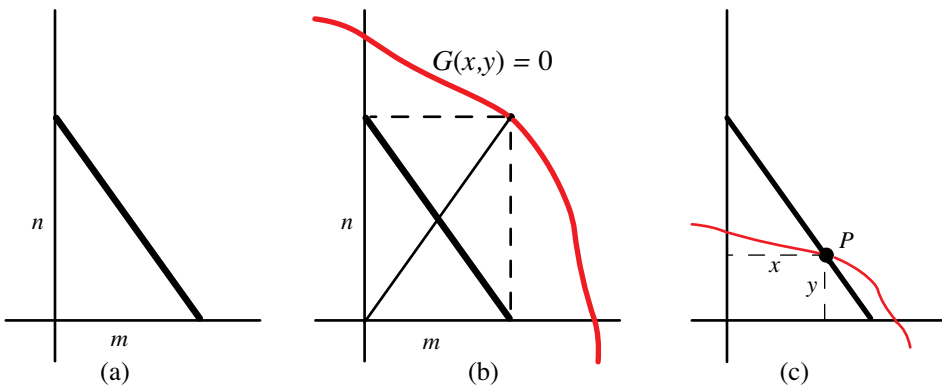


Figure 9.15: (a) Trammel with intercepts m and n. (b) Point (m, n) is on the governing curve $G(x, y) = 0$. (c) A point P that divides the trammel into segments of prescribed ratio λ/μ with $\lambda + \mu = 1$ lies on the path given by $G(x/\lambda, y/\mu) = 0$.

Instead of m^2+n^2 being constant, we assume that the intercepts (m, n) satisfy a more general relation, say

$$G(x, y) = 0, \tag{9.14}$$

where G is a function we call the *governor* of the trammel. If the intercepts m and n satisfy $G(m,n)=0$ we refer to the portion of the line (9.13) in a given quadrant as a *flexible trammel governed by G*. The length of the trammel is equal to that of the radial segment from the origin to (m,n), where $G(m,n)=0$. The segment and the trammel are the diagonals of a rectangle with opposite vertices at the origin and at (m,n) as shown in Figure 9.15b. This extends Figure 9.2c for a standard trammel.

When $G(m,n)=m^2+n^2-1$, the trammel has constant length 1. As this standard trammel slides along the axes, a point P on it traces an ellipse. Now we seek the path traced by a point P on a flexible trammel as the trammel slides along the axes. We treat only the simple case when *P divides the trammel into two pieces whose lengths are in constant ratio.* It can be realized physically by an elastic string with a knot on it.

Theorem 9.5. *Consider a flexible trammel governed by a function G. A point P that divides the trammel into two pieces whose lengths are in a ratio λ/μ, where λ and μ are nonzero constants with $\lambda+\mu=1$, lies on the path given by*

$$G(\frac{x}{\lambda},\frac{y}{\mu})=0. \tag{9.15}$$

Proof. Point P has coordinates $(\lambda m,\mu n)$, and because of (9.14) it lies on the path given by (9.15).

Definition. The graph of (9.15) is called the *trace* of P, denoted by $\tau(\lambda)$.

Theorem 9.5 shows how easy it is to determine the trace of P once the governing function of the trammel is known.

Because $\lambda+\mu=1$ the curves $\tau(\lambda)$ form a one-parameter family of curves determined by the governor G and the choice of λ. The trace $\tau(\lambda)$ is obtained by dilating the graph of G horizontally by a factor λ and vertically by a factor $\mu=1-\lambda$.

If both m and n are positive, the trammel lies in the first quadrant. The restriction $\lambda+\mu=1$ keeps $P=(\lambda m,\mu n)$ in the first quadrant between the ends of the trammel if both λ and μ are positive. But if $\lambda>1$ then $\mu=1-\lambda$ is negative and P lies in the fourth quadrant. Similarly, if $\mu>1$ then $\lambda=1-\mu$ is negative and P lies in the second quadrant.

9.10 TANGENCY OF A FLEXIBLE TRAMMEL AND ITS TRACE

In Section 9.5 we found the point of tangency of a standard trammel and the ellipse traced by a point on the trammel. Now we do the same for the curve in (9.15) traced by a point P that divides a flexible trammel into pieces of ratio λ/μ. For simplicity, assume that the governing equation $G(x,y)=0$ in (9.14) can be solved for y in terms of x, giving

$$y=g(x). \tag{9.16}$$

Then the equation of the trace of a point $P=(\lambda m,\mu n)$ on the trammel is

$$y=\mu g(\frac{x}{\lambda}). \tag{9.17}$$

The curve in (9.17) has slope

$$\frac{dy}{dx} = \frac{\mu}{\lambda} g'(\frac{x}{\lambda})$$

at each of its points, while the trammel has slope $-n/m$ at each of its points. Equating the slopes at the point P of intersection we find

$$-\frac{n}{m} = \frac{\mu}{\lambda}\frac{dn}{dm}, \tag{9.18}$$

where $dn/dm = g'(m)$.

Equation (9.18) determines the ratio n/m implicitly in terms of μ, λ, and G, and we denote by ρ the ratio n/m so defined. The ratio ρ is the tangent of the critical angle θ that the trammel makes with the x axis when it is tangent to the trace $\tau(\lambda)$. Consequently, (9.18) gives the following extension of Theorem 9.1:

Theorem 9.6. *The trammel is tangent to the trace when its angle of inclination θ with the x axis satisfies*

$$\tan\theta = \rho. \tag{9.19}$$

To show how (9.19) can be used to determine the slope of the trammel when it is tangent to the trace $\tau(\lambda)$ we consider the following example.

Example 1: $G(x, y) = (\frac{x}{A})^k + (\frac{y}{B})^k - 1$ (kth power ellipse).

The trace in this example is given by

$$(\frac{x}{\lambda A})^k + (\frac{y}{\mu B})^k = 1.$$

The equation $G(m, n) = 0$ becomes $m^k B^k + n^k A^k = A^k B^k$, which yields

$$\frac{dn}{dm} = -\frac{B^k m^{k-1}}{A^k n^{k-1}}. \tag{9.20}$$

Equation (9.18) implies $(n/m)^k = (B/A)^k(\mu/\lambda)$, and when this is solved for $\rho = n/m$ we obtain

$$\tan\theta = \rho = \frac{B}{A}(\frac{\mu}{\lambda})^{\frac{1}{k}}. \tag{9.21}$$

When $k = 2$ and $A = B$ this agrees with the formula obtained earlier in (9.8).

The case $k = 1$ is also of interest because the governor is a straight line $x/A + y/B = 1$ and every trace is also a straight line $x/(\lambda A) + y/(\mu B) = 1$. In this case (9.21) becomes

$$\tan\theta = \frac{n}{m} = \frac{B}{A}(\frac{\mu}{\lambda}). \tag{9.22}$$

The trammel and trace are straight lines that intersect at exactly one point unless they have the same slope, in which case the trammel and trace coincide. From (9.22) we see that this happens if and only if $\mu/\lambda = An/Bm$.

9.11 ENVELOPE OF A TRAMMEL AND OF ITS FAMILY OF TRACES

As the angle α of inclination of a trammel varies it generates a one-parameter family of lines that has an envelope ε touching all of them. For a given position of the trammel select the point of tangency with ε. That point divides the trammel into pieces having a ratio λ/μ with $\lambda + \mu = 1$. Point $P = (\lambda m, \mu n)$ traces a curve $\tau(\lambda)$ given by (9.17). The trace cannot cross ε in a neighborhood of P, otherwise the trammel would cross its envelope at P. Therefore the trace is also tangent to ε at P. If ε exists we can use (9.18) and (9.13) to determine parametric equations for it. First, we have

$$\frac{x}{m} = \lambda = \frac{\lambda}{\lambda + \mu} = \frac{1}{1 + \frac{\mu}{\lambda}},$$

which in view of (9.18) and (9.13) gives us the pair of equations

$$\frac{x}{m} = \frac{1}{1 - \frac{n}{m}\frac{dm}{dn}}, \quad \frac{y}{n} = 1 - \frac{x}{m}. \tag{9.23}$$

This gives us the following:

Theorem 9.7. *The envelope of a moving flexible trammel is also the envelope of the family of traces of the trammel, as described by* (9.23).

The second equation in (9.23) can be written as

$$\frac{y}{n} = \frac{1}{1 - \frac{m}{n}\frac{dn}{dm}}. \tag{9.24}$$

Because m and n satisfy the governing condition $G(m, n) = 0$, relations (9.23) and (9.24) serve as parametric equations for the envelope ε in terms of one of the parameters m or n.

In general the derivatives dm/dn and dn/dm can be obtained by differentiating the implicit equation $G(m, n) = 0$, as is done for the following examples.

Example 1: $G(x, y) = (\frac{x}{A})^k + (\frac{y}{B})^k - 1$ (kth power ellipse).

In this case we use (9.20) for dn/dm, and the parametric equations (9.23) and (9.24) yield $x/A = (m/A)^3$, $y/B = (n/B)^3$. Eliminating the parameters m, n we find that the envelope of the family of kth power ellipses is a generalized astroid given by

$$(\frac{x}{A})^{\frac{k}{k+1}} + (\frac{y}{B})^{\frac{k}{k+1}} = 1. \tag{9.25}$$

Figure 9.16 shows some special cases. The dashed line in Figure 9.16a shows the case $k = 1$ with $A = 2$, $B = 3$. In this case the envelope ε is the parabola described by

$$\sqrt{x/2} + \sqrt{y/3} = 1.$$

The symmetric case $A = B$ (Figure 9.16d) yields a known parabolic envelope [70, p. 76]. Figure 9.16b shows the case $k = 2$ (an ordinary ellipse with $A = 2, B = 3$) whose envelope is an asymmetric astroid.

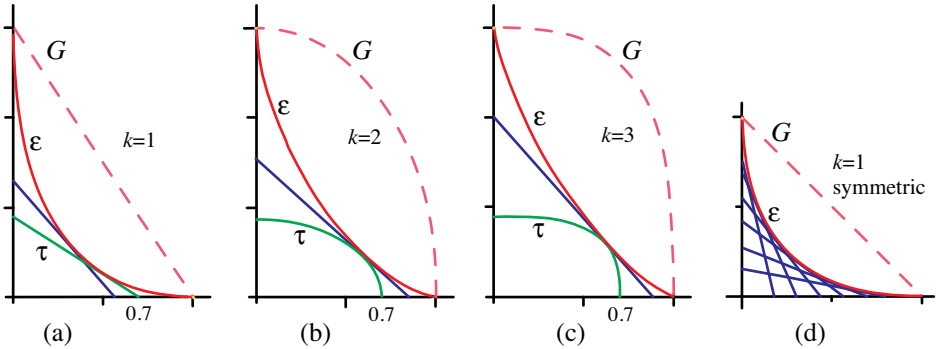


Figure 9.16: Special cases of Example 1 showing the governor G (dashed), and the trammel tangent to the envelope ε. In (a)-(c), the trace curve $\tau(0.7)$ is shown. In (d) the governor is $x + y = 1$, shown in the first quadrant. When the governor is in the second or fourth quadrant, the parabolic envelope continues in the first quadrant with the line $y = x$ as its axis of symmetry.

Figures 9.16b-c show the trammel tangent to ε, and to the trace curve $\tau(0.7)$.

Example 2: $G(x, y) = y - x^k$ (kth power function).

In this case the trace is another kth power function given by

$$y = \mu(\frac{x}{\lambda})^k.$$

Figure 9.17a shows the case $k = 1$, in which the governor is the line $y = x$, the trace curve is the line $y = \mu x/\lambda$, which is never tangent to the trammel, and there is no envelope.

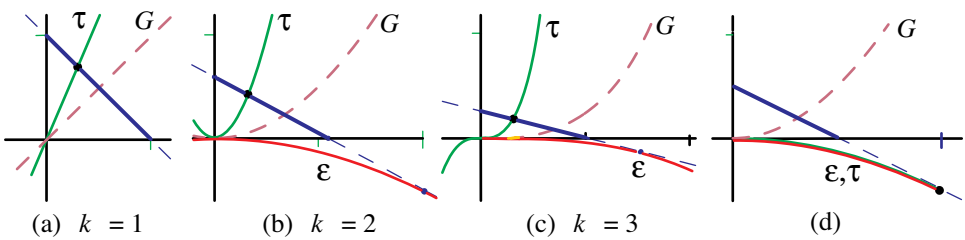


Figure 9.17: Special cases of Example 2 with positive k, and $\lambda = 0.3$ in (a)-(c).

If $k \neq 1$, the envelope is yet another kth power function given by

$$y = \frac{(\frac{k-1}{k})^k}{1-k} x^k.$$

Moreover, in this case the envelope touches all the trace curves only at the origin. The line through the trammel is always tangent to the envelope in the fourth quadrant.

Figure 9.17d shows a special feature that holds for any exponent k. The trammel and the trace have the same slope at their point of intersection when (9.19) is satisfied. If $n = m^k$ then $dn/dm = km^{k-1} = kn/m$, and (9.19) implies

$$\frac{n}{m}(\frac{\mu k}{\lambda} + 1) = 0. \tag{9.26}$$

Relation (9.26) is satisfied if n/m=0 (when the trammel is horizontal) but (9.26) also holds for a nonhorizontal trammel when the point of subdivision $(\lambda m, \mu n)$ on the trammel satisifes $\lambda/\mu = -k$, or $\lambda = k/(k-1), k \neq 1$. For this choice of λ, the trace $\tau(\lambda)$ coincides with the envelope, and a line through the trammel is tangent to both in the fourth quadrant as shown by the example in Figure 9.17d.

Figure 9.18 shows special cases of Example 2 when the exponent k is negative. In Figure 9.18a, b, c we have $k = -1, k = -2, k = -3$, respectively. The exceptional case (9.26) is shown in Figure 9.18d. For the special choice of ratio $\lambda/\mu = -k$, the trace $\tau(\lambda)$ coincides with the envelope and the trammel is tangent to both in the first quadrant as indicated in Figure 9.18d.

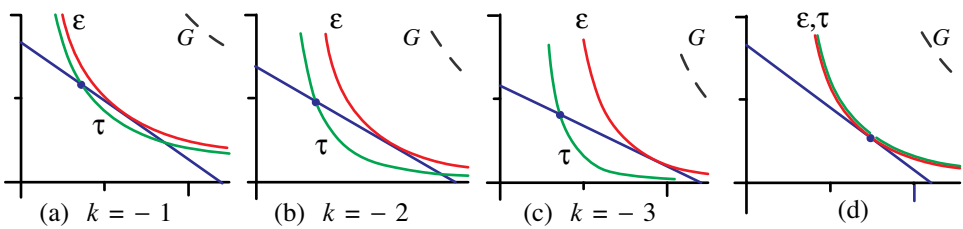


Figure 9.18: Special cases of Example 2 with negative k, and $\lambda = 0.3$ in (a)-(c).

The governing curve, each trace, and the envelope in Figure 9.18a are rectangular hyperbolas with the axes as asymptotes. The next example treats kth power governing hyperbolas written in standard form.

Example 3: $G(x, y) = (x/A)^k - (y/B)^k - 1$ (kth power hyperbola, $k \neq 0$).

By Theorem 9.5 the trace is also a generalized kth power hyperbola given by

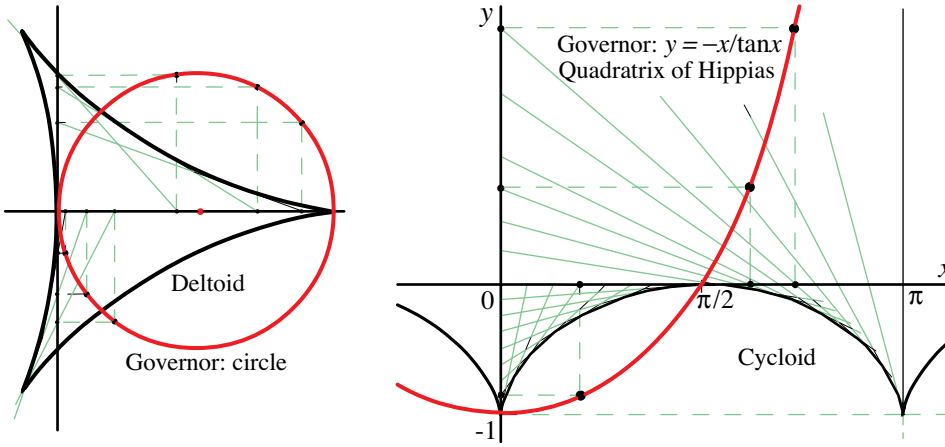
$$(x/\lambda A)^k - (y/\mu B)^k = 1.$$

If $k \neq -1$, the envelope has cartesian equation resembling that in (9.25) with a difference instead of a sum:

$$(\frac{x}{A})^{\frac{k}{k+1}} - (\frac{y}{B})^{\frac{k}{k+1}} = 1.$$

9.12 APPLICATION: GRAPHIC CONSTRUCTION OF ENVELOPES AND GOVERNORS AS A CLASSROOM ACTIVITY

The foregoing examples suggest a simple and engaging educational activity that can be performed on graph paper. It reveals the qualitative shape of the envelope of a family of tangent lines without the use of equations for the envelope.



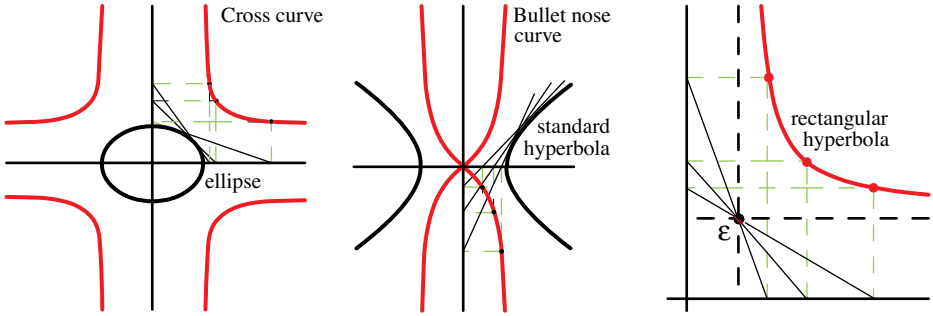
(a) Deltoid as envelope (b) Cycloid as envelope

Figure 9.19: (a) A circular governor tangent to the y axis produces a deltoid as envelope. (b) Cycloid as envelope for the quadratrix of Hippias as governor. Different branches of the quadratrix produce copies of the cycloid.

Start with a straight line as governor G, as shown by the dashed line in Figure 9.16a. Choose a point (m, n) on G, and drop perpendiculars to the coordinate axes to obtain the projections $(m, 0)$ and $(0, n)$ on the axes. The line drawn through the projected points has intercepts m and n and is tangent to the corresponding envelope ε. By choosing several points on G and repeating this process, the positions of the tangent lines gradually reveal the shape of the envelope.

For a more general governor G the corresponding envelope ε can be constructed in exactly the same manner. The following examples are of special interest.
1. If G is a straight line, ε is a parabola (Figures 9.16a and d).
2. If G is a circle centered at the origin (as in Figure 9.16b), ε is an astroid.
3. If G is a circle tangent to the y axis with center on the x axis as shown in Figure 9.19a, the envelope is a *deltoid* (see Section 2.4).
4. If G is the quadratrix of Hippias, known since 430 B.C., (see [70, p. 204]), ε is a *cycloid* (Figure 9.19b). This connection between the quadratrix of Hippias and the cycloid is completely unexpected.

We can also start with a prescribed curve as envelope and determine the governor leading to it by simply reversing the steps described above, as illustrated in Figures 9.20 and 9.21 for the following examples.
5. When the envelope ε is an ellipse centered at the origin as shown in Figure 9.20a, the governor is a *cross curve* (see [70, p. 203]).
6. When the envelope ε is a hyperbola centered at the origin as shown in Figure 9.20b, the governor is a *bullet nose curve* (see [70, p. 203]).
7. When the envelope ε is a single point, the governor is a rectangular hyperbola with asymptotes intersecting at that point (Figure 9.20c).



(a) Cross curve as governor (b) Bullet nose curve as governor (c) Single point as envelope

Figure 9.20: (a) An ellipse as envelope arising from a cross curve. (b) A hyperbola as envelope arising from a bullet nose curve. (c) A single point as envelope of a shifted hyperbola.

8. When the envelope ε is a circle tangent to the coordinate axes as in Figure 9.21a, the governor is a curve that resembles a *bicorn* or "Napoleon's hat."
9. When the envelope ε is the cardioid shown in Figure 9.21b, the governor is a curve whose cartesian equation is $y = (9x^2 - 27)/(27 - x^2)$.

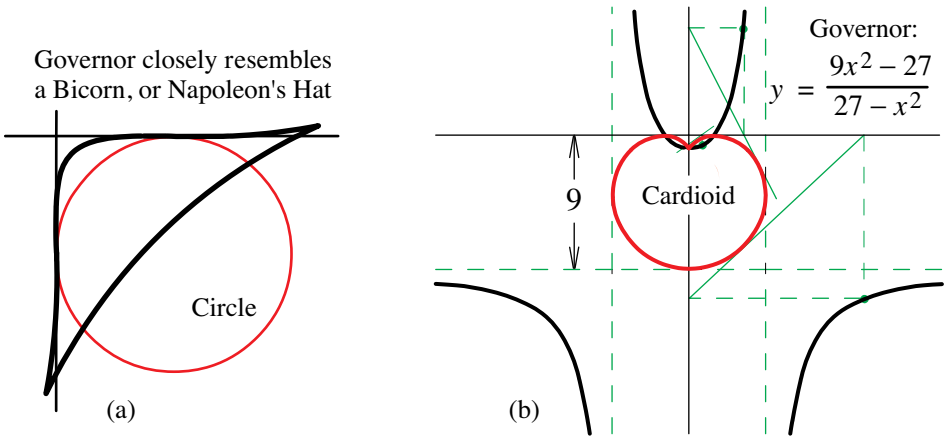


Figure 9.21: In (a) the envelope is a circle. In (b) the envelope is a cardioid.

Also, if the trammel is realized by an elastic string tangent to the envelope, a knot at its midpoint will trace a curve similar to the governor. A knot elsewhere will trace a dilated copy of the governor (Theorem 9.5).

9.13 REMARKS CONCERNING HOLDITCH'S THEOREM

In 1858 the Reverend Hamnet Holditch, president of Caius College in Cambridge, published a one-page note [49] announcing the following surprising result, which became known as *Holditch's theorem*:

If a chord of a closed plane curve be divided into two parts of constant lengths a and b, the difference between the areas bounded by the closed curve and by the locus of the dividing point, will be πab.

An example illustrating Holditch's theorem is shown in Figure 9.22. The outer curve C is the given closed curve, and chord AB is divided by point D. The inner curve E is the envelope of the moving chord AB, which is always tangent to E. The locus of the division points D is the trace T, shown intermediate to E and C.

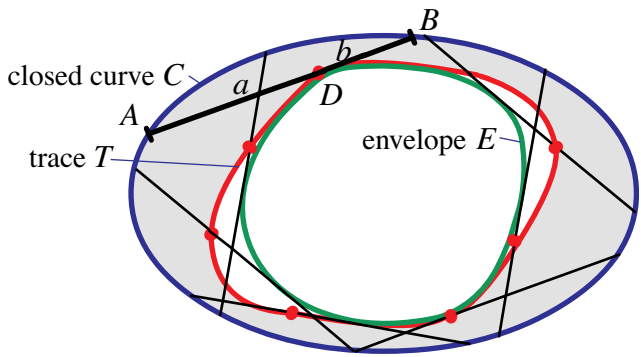


Figure 9.22: Illustrating Holditch's theorem. The chord AB of length $a + b$ moves once around the interior of the closed curve C. The trace T is the locus of the division points D. The shaded region between T and C has area πab.

Holditch's theorem states that the area of the region between C and T is πab. This is surprising because it depends only on the product of the lengths a and b and not on the size or shape of curve C.

Holditch's original note ignores several assumptions that are needed to validate his proof for a general closed curve C. His argument implicitly assumes that C is convex (his drawing shows C as a circle) and that the chord can actually navigate around it in such a way that the dividing point traces out another closed curve inside the curve. A discussion of these unstated assumptions has been given in several papers, notably [30] and [34], which together with [32] also provide various extensions of Holditch's theorem.

We shall use sweeping tangents to derive a bipartite sweeping formula (9.29) that not only implies Holditch's theorem but extends the generalization in [32] as well. In our general treatment a tangent segment of *variable length* $a + b$ moves along a *space curve* E, and its endpoints A and B trace space curves that are not necessarily closed. Our formula can be used to calculate areas of classical regions, such as the elliptical sector in Theorem 9.2, the region bounded by a cardioid, or the region

bounded by the limaçon of Pascal. Although our general formula extends Holditch's theorem, we have not found any interesting applications to areas of classical curves that cannot be treated more easily by standard calculus integration.

A bipartite sweeping formula.

Start with a space curve E that will play the role of the envelope of the chords in Holditch's treatment. Imagine a moving coordinate axis with its origin initially tangent to E. The coordinate axis consists of two infinite rays, one positive and one negative, joined at the origin. If we allow the axis to move tangent to E, each ray sweeps out one sheet of a developable surface. The two sheets meet along E, and their union is the tangent developable surface of E. The curve E is called the edge of regression of the surface (see [67; p. 71]). Now consider the developable surface swept by a tangent segment of variable length $a+b$ moving along E, requiring only that a and b be proportional. During this motion, the two endpoints A and B of the tangent segment trace portions of two curves (instead of Holditch's closed curve), one on each sheet of the developable surface. Formula (9.29) relates the areas of the regions swept by the two pieces of lengths a and b, hence the name *bipartite formula*. The formula is also meaningful if the dividing point D is not between the two endpoints, provided it lies on the line determined by the moving tangent. We emphasize again that our treatment does not require that the endpoints A and B move along closed curves.

Let φ denote the angle that the positive axis sweeps out as it moves from a fixed tangent line to a general position. Consider three points A, B, D on the same ray of the axis with respective coordinates α, β, δ that vary continuously with φ. When $0 < \beta < \delta < \alpha$, D lies between B and A and divides BA into segments of lengths $(\delta - \beta)$ and $(\alpha - \delta)$, as shown in Figure 9.23.

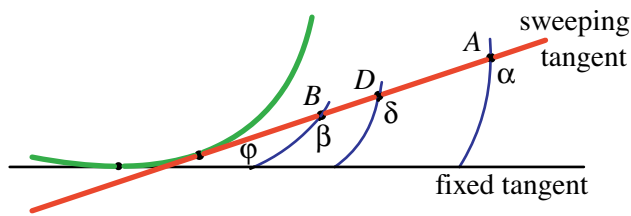


Figure 9.23: When $0 < \beta < \delta < \alpha$, D divides BA into segments of lengths $(\delta - \beta)$ and $(\alpha - \delta)$.

We consider the integrals

$$H_A = \frac{1}{2}\int_0^\theta (\alpha^2 - \delta^2)\, d\varphi, \text{ and } H_B = \frac{1}{2}\int_0^\theta (\delta^2 - \beta^2)\, d\varphi. \tag{9.27}$$

These integrals are meaningful for any configuration of points A, B, D, but when D lies between B and A as in Figure 9.23, H_A represents the area of a region swept by segment DA between the curves traced by α and δ, and H_B represents the area swept by segment BD between the curves traced by δ and β.

Now introduce the functions

$$a(\varphi) = \alpha - \delta, \text{ and } b(\varphi) = \delta - \beta,$$

which can be positive or negative. In the configuration shown in Figure 9.23, both $a(\varphi)$ and $b(\varphi)$ are positive because they are the lengths of segments DA and BD, respectively.

In general, $\alpha = a(\varphi) + \delta$, $\beta = \delta - b(\varphi)$, and the integrals for H_A and H_B become

$$H_A = \frac{1}{2}\int_0^\theta \left((a(\varphi)+\delta)^2 - \delta^2\right) d\varphi = \frac{1}{2}\int_0^\theta a^2(\varphi)d\varphi + \int_0^\theta a(\varphi)\delta d\varphi,$$

$$H_B = \frac{1}{2}\int_0^\theta (\delta^2 - (\delta - b(\varphi))^2)d\varphi = \frac{-1}{2}\int_0^\theta b^2(\varphi)d\varphi + \int_0^\theta b(\varphi)\delta d\varphi.$$

At this point we add the restriction that $a(\varphi)$ and $b(\varphi)$ are proportional. Specifically, assume that

$$qa(\varphi) = pb(\varphi), \tag{9.28}$$

where p and q are real constants with $p + q = 1$. In Figure 9.23, this means that D divides BA into two pieces whose lengths have constant ratio. The linear combination $qH_A - pH_B$ eliminates the integrals involving δ and gives us

$$qH_A - pH_B = \frac{1}{2}\int_0^\theta \left(qa^2(\varphi) + pb^2(\varphi)\right) d\varphi.$$

But $qa^2(\varphi) + pb^2(\varphi) = a(\varphi)pb(\varphi) + b(\varphi)qa(\varphi) = a(\varphi)b(\varphi)$ by (9.28), and we find the *bipartite sweeping formula*:

$$qH_A - pH_B = \frac{1}{2}\int_0^\theta a(\varphi)b(\varphi)\, d\varphi. \tag{9.29}$$

This is a property of the integrals in (9.27). When properly interpreted, it provides a relation between areas, as shown by examples below.

When A and B move once around the same plane closed curve, (9.29) reduces to a result of Chakerian and Goodey [32].

Holditch's theorem as a special case.

To show that Holditch's theorem is a consequence of (9.29), take $a(\varphi) = a$ and $b(\varphi) = b$, where a and b are constant subject to the proportionality relation $qa = pb$, where $p + q = 1$. Then (9.29) becomes

$$qH_A - pH_B = \frac{\theta}{2}ab. \tag{9.30}$$

Now assume that A and B move once around the same closed curve C. Then $\theta = 2\pi$, and (9.30) becomes

$$qH_A - pH_B = \pi ab. \tag{9.31}$$

In this case, the integral $H_A = \frac{1}{2}\int_0^{2\pi}(\alpha^2-\delta^2)\,d\varphi$ is positive and represents the area of the ring swept by AD, which lies between C and the curve traced by D. On the other hand, B traces the same closed curve as A so we have

$$H_B = \frac{1}{2}\int_0^{2\pi}(\delta^2-\beta^2)\,d\varphi = \frac{1}{2}\int_0^{2\pi}(\delta^2-\alpha^2)\,d\varphi = -H_A.$$

Therefore $qH_A - pH_B = (q+p)H_A = H_A$ because $q+p=1$, so (9.31) yields Holditch's theorem:

$$H_A = \pi ab.$$

Alternative form of the bipartite formula.

We can express (9.29) in another form by introducing

$$c(\varphi) = a(\varphi) + b(\varphi) = \alpha - \beta,$$

which, for the configuration in Figure 9.23, represents the length of segment BA. The relation $p+q=1$, together with (9.28), gives us

$$b(\varphi) = pb(\varphi) + qb(\varphi) = q(a(\varphi)+b(\varphi)) = qc(\varphi),$$

and

$$a(\varphi) = pa(\varphi) + qa(\varphi) = p(a(\varphi)+b(\varphi)) = pc(\varphi),$$

hence

$$a(\varphi)b(\varphi) = pq\,c^2(\varphi).$$

Therefore the bipartite sweeping formula (9.29) can be written in the alternative form

$$qH_A - pH_B = \frac{pq}{2}\int_0^{\theta} c^2(\varphi)\,d\varphi. \tag{9.32}$$

NOTES ON CHAPTER 9

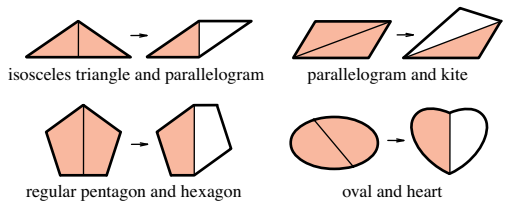
Except for Section 9.13 on Holditch's theorem, the material in this chapter originally appeared in [23]. The results in Section 9.13 have not been previously published.

Chapter 10

ISOPERIMETRIC AND ISOPARAMETRIC PROBLEMS

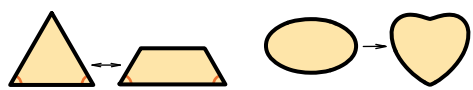
These problems can be easily solved by the methods developed in this chapter. The reader may wish to try solving them before reading the chapter.

Four pairs of isoparametric plane regions (having equal areas and equal perimeters), are shown. Each pair is an example of special dissections, in which a chord bisects the first region, and one of the two pieces is flipped to produce a region isoparametric to the first. In these examples, the boundary of the first region is converted onto the boundary of the second.



A dissection that converts boundaries onto one another is called complete.

Find a complete dissection, without flipping, that converts the isosceles triangle below onto the isoparametric isosceles trapezoid shown. Do the same for the oval and heart shapes.



CONTENTS

PART 1: ISOPARAMETRIC REGIONS

PART 2: DISSECTIONS OF ISOPARAMETRIC REGIONS

10

Traditional isoperimetric problems ask for the region of maximal area among all plane regions having equal perimeters. Part 1 of this chapter deals instead with plane regions that have equal perimeters and equal areas. We call such regions isoparametric and introduce the isoparametric problem: find incongruent isoparametric regions of specified shapes. Surprisingly, incredibly many solutions exist, and the problem opens a broad new field of research. Simple examples include a square and two different circular sectors, and a rectangle and two different isosceles triangles. The general problem is analyzed with the help of the contour ratio, a replacement of the classical isoperimetric quotient. Two contours can be scaled to become isoparametric if and only if they have the same contour ratio. The problem becomes more interesting and more difficult when it is extended to rings, that is, regions between two similar simple closed curves. For example, although a square and a circular disk are never isoparametric, a square ring and a circular ring can be isoparametric, unless the hole in the circular ring is too small to make a significant contribution to the total perimeter and area.

Part 2 introduces a new development in classical dissection problems: *complete dissections* that convert regions and their boundaries into one another. This requires isoparametric regions as introduced in Part 1. To the best of our knowledge, complete dissections have not been previously treated.

PART 1: ISOPARAMETRIC REGIONS

10.1 INTRODUCTION

Two incongruent solids with remarkable properties are shown in Figure 10.1. One is a slice of a solid hemispherical shell with inner radius r and outer radius R cut by a plane parallel to the equator and at distance $h < r$ from the equator. The

other is a cylindrical shell with the same radii and altitude h. The surface of each solid consists of four components: an upper circular ring, a lower circular ring, an outer lateral surface, and an inner lateral surface. The two solids share the following properties:

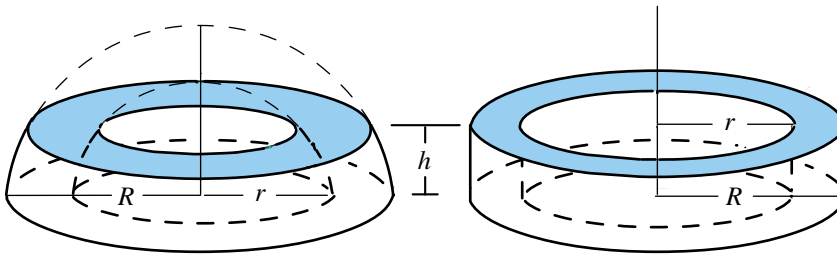


Figure 10.1: Incongruent solids sharing the properties (a) through (e).

(a) The solids have equal volumes.

(b) The solids have equal total surface area.

(c) Every plane parallel to the equator cuts both solids in cross sections of equal area.

(d) The two inner lateral surface areas are equal.

(e) The two outer lateral surface areas are equal.

These two solids were introduced in Chapter 5, where it was shown that the properties (a)–(e) are shared by an entire family of incongruent solids, each of which has polygonal rings as cross sections, rings formed by similar polygons circumscribing the inner and outer circular cross sections of a spherical shell. These solids are part of a more general family with polygonal rings as cross sections (described in Section 10.8) that is even more remarkable because it satisfies the five properties and a sixth property not shared by the solids in Figure 10.1:

(f) Every plane parallel to the equator cuts both solids in cross-sectional rings whose inner perimeters are equal and whose outer perimeters are equal.

The last property implies that the two cross-sectional rings also have the same total perimeter. This observation motivated the present chapter, which is concerned with plane regions having equal areas and equal perimeters. Because two global parameters (area and perimeter) are to be equal, we call such regions *isoparametric*. The first problem we posed was

How can we construct incongruent isoparametric plane regions?

With no further restrictions on the regions, it is easy to produce examples at will, as shown by the regions in Figure 10.2. A chord divides each region into two pieces. One piece is flipped to produce an incongruent isoparametric region. It is clear that infinitely many such regions can be produced in this way by cutting and flipping a piece from any region, unless it is a circular disk, in which case flipping one of the pieces results in a congruent disk.

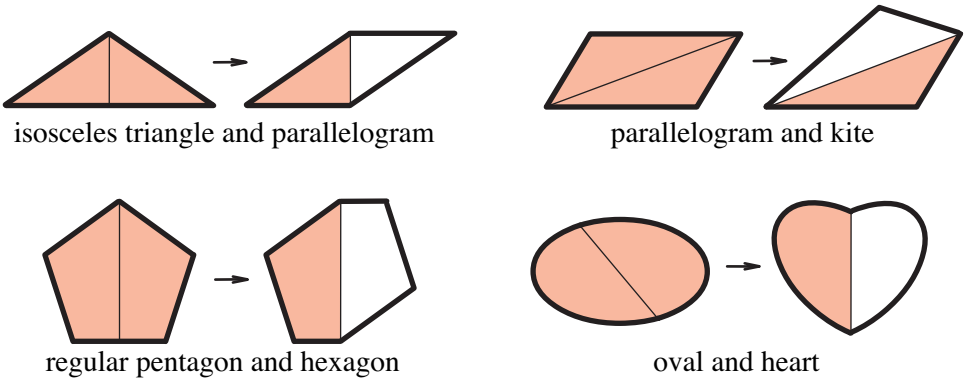


Figure 10.2: Incongruent isoparametric plane regions formed by cutting and flipping.

Traditional isoperimetric problems compare different plane regions having equal perimeters and ask for the region of maximal area. It is known [35], [61] that among all regions with a given perimeter, the circle encloses the largest area. This follows from the isoperimetric inequality,

$$\frac{p^2}{4s} \geq \pi, \tag{10.1}$$

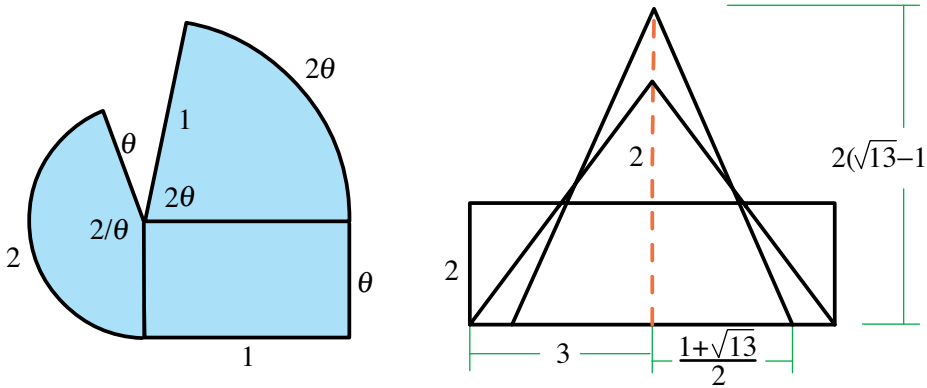
which relates the perimeter p and area s of a planar region bounded by a simple closed curve. Equality holds in (10.1) only for the circle.

Isoperimetric problems have been a source of important mathematical ideas and techniques since classical antiquity. A result arising from mythology is Dido's problem: in the half-plane bounded by a line, find a curvilinear arc of prescribed length with its extremities on the line and enclosing the maximum area. The solution, a semicircle whose diameter is on the given line, is obtained by reflecting the curve in the line and invoking the isoperimetric property of the circle. Archimedes treats a three-dimensional analog in Proposition 9 of his *Sphere and Cylinder II*, which states that of all spherical segments having equal spherical surface area, the hemisphere has the greatest volume. Today, isoperimetric problems and their extensions are alive and well. They continue to nourish mathematical imagination, as evidenced by a recent proof of the double bubble conjecture [50]. Interesting historical perspective on isoperimetric problems is given in [35], which also describes their relation to a host of other maximum-minimum problems dealt with by a method called the calculus of variations.

This chapter treats a different type of problem: find incongruent plane regions that have equal perimeters and equal areas. Hence the new name: isoparametric problem. The problem becomes more interesting, and more difficult, if we seek incongruent isoparametric regions of specified shapes. It cannot be solved if one of the regions is a circular disk because of the isoperimetric inequality. Also, it cannot be solved for two regular polygons with different numbers of sides (see Section 10.2).

The authors were pleasantly surprised to discover numerous cases where it can

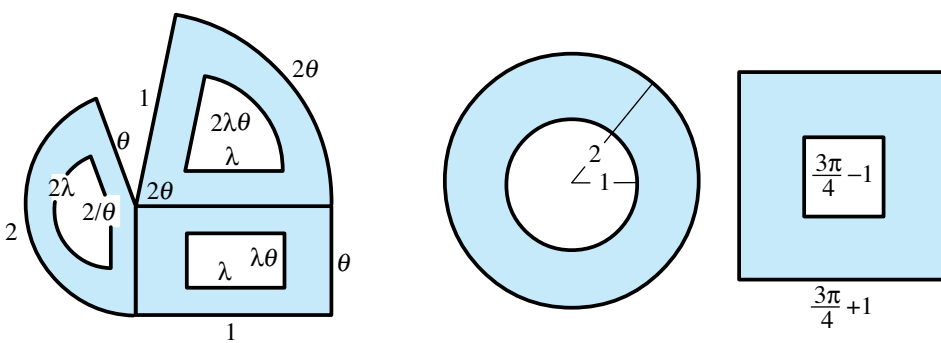
be solved. Figure 10.3a shows three incongruent isoparametric regions: two circular sectors and a rectangle, each with area θ and perimeter $2 + 2\theta$. Because θ is arbitrary, this provides an infinite family of such triples. When $\theta = 1$ the rectangle is a square and the two sectors are congruent. Figure 10.3b shows three more examples: two isosceles triangles and a rectangle, each with area 12 and perimeter 16.



(a) Two circular sectors and a rectangle (b) Two isosceles triangles and a rectangle

Figure 10.3: Examples of incongruent isoparametric regions.

Such examples show that the general problem of finding incongruent isoparametric regions of specified shapes opens a door to many possibilities worth exploring. Section 10.3 gives a systematic treatment. Section 10.4 treats the same type of problem for rings bounded by two similar closed curves. Introducing "holes" makes the problem more interesting and allows more possibilities. For example, Figure 10.4a shows three rings formed from the sectors and rectangle in Figure 10.3a. The



(a) Sectorial rings and rectangular rings (b) Circular ring and square ring

Figure 10.4: Examples of incongruent isoparametric rings.

holes are obtained by shrinking each figure by the same size factor $\lambda\,(<1)$. For any size factor λ, each of the rings has area $(1-\lambda^2)\theta$ and total perimeter $(1+\lambda)(2+\theta)$,

so they are isoparametric. In Figure 10.4b a circular ring and a square ring have the same area 3π and the same total perimeter 6π. For the circular ring the size factor λ (ratio of inner to outer radius) is $1/2$. But if $\lambda < 1/9$, it turns out that there is no isoparametric square ring. Sections 10.5 and 10.6 explain why there is a restriction on λ in Figure 10.4b but not in Figure 10.4a. Section 10.7 discusses isoparametric rings that have property (f) mentioned earlier.

10.2 CONTOUR RATIOS

In this chapter, a contour is a plane region that has associated with it a perimeter p and an area s. For a given region, the ratio $Q = 4\pi s/p^2$ is called the isoperimetric quotient. The isoperimetric inequality states that $Q \leq 1$ for regions bounded by simple closed curves, with $Q = 1$ only for the circle. Some properties of Q are given in [52]. For our purposes, it is more useful to study the quotient

$$\kappa = \frac{p^2}{4s}, \tag{10.2}$$

which we call the *contour ratio*. It has the pleasant feature that $\kappa = \pi$ for a circle and $\kappa = 4$ for a square. Isoparametric contours have the same contour ratio. A regular n-gon has contour ratio $\kappa = n\tan(\pi/n)$, a decreasing function of n that approaches π as $n \to \infty$. That is why regular polygons with different numbers of sides cannot be isoparametric.

For all contours with $s = 1$, the contour ratio is the square of the semiperimeter. Qualitatively, the contour ratio indicates the dominance of the semiperimeter over the square root of the area. Similar contours have the same contour ratio because the scaling factor cancels in (10.2).

Figure 10.5 shows various shapes arranged by contour ratios. The size of a region plays no role. All circles are located at π, and all squares at 4.

Figure 10.5 also provides a spectrum of contour ratios for families of shapes. Regular polygons serve as discrete bench marks. For example, an equilateral triangle has contour ratio $3\sqrt{3} = 5.1961$ .... All other triangles have larger contour ratios, so their images are distributed continuously to the right of the equilateral triangle. The square, another bench mark, has the smallest contour ratio of all quadrilaterals. More generally, for each n the images of all n-gons lie to the right of the regular n-gon, which has the smallest contour ratio among all n-gons.

An isosceles right triangle is also a bench mark. Its contour ratio is $3 + 2\sqrt{2} = 5.8284...$, which is the smallest κ that occurs among all right triangles. In particular, any right triangle with integer sides (a Pythagorean triangle) has a larger κ. The Pythagorean 3 :4 :5 triangle has $\kappa = 6$, and the Pythagorean 119 :120 :169 triangle, which is nearly isosceles, has $\kappa = 5.8285...$. We offer as a challenge to the reader to show that Pythagorean right triangles exist with κ arbitrarily close to $3 + 2\sqrt{2}$; thus, no Pythagorean triangle exists with smallest κ.

On the other hand, two regions can have the same contour ratio even though their shapes are quite different. In Figure 10.3a the rectangle has a different shape than the two sectors, but each has area θ and perimeter $2\theta + 2$, so their contour ratios are equal.

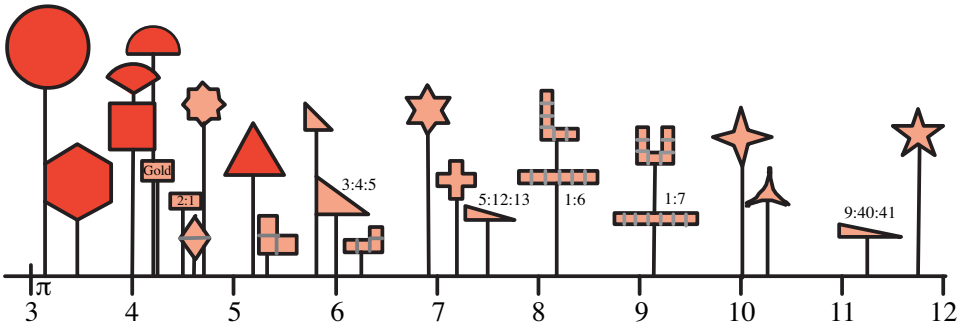


Figure 10.5: A spectrum of contour ratios of various shapes. Relative sizes are irrelevant.

The pentagram in Figure 10.5 has contour ratio $10\sqrt{3-\phi} = 11.7557...$, where $\phi = (\sqrt{5}+1)/2$ is the golden ratio. A special long thin rectangle of some base $b \approx 9.6521$ and height 1 has the same contour ratio, even though its shape has no resemblance to a pentagram. Although the perimeter and area of the rectangle are not equal to those of the pentagram, there is a similar rectangle with exactly the same area and the same perimeter as the pentagram, as revealed by the following theorem.

Theorem 10.1. (Isoparametric Contour Theorem) *Two contours can be scaled to become isoparametric if and only if they have the same contour ratio.*

Proof. If two contours can be scaled to become isoparametric, then the scaled contours have the same contour ratio, hence so do the original contours because κ is invariant under scaling. Conversely, assume that contours 1 and 2 have equal contour ratios, $p_1^2/(4s_1) = p_2^2/(4s_2)$. Then $s_2/s_1 = (p_2/p_1)^2$. If we scale contour 1 by the scaling factor $t = p_2/p_1$ we obtain a similar contour with perimeter $tp_1 = p_2$ and area $t^2 s_1 = s_2$. This scaled copy of contour 1 is isoparametric to contour 2. In the same way, if we scale contour 2 by the scaling factor p_1/p_2 the scaled copy will be isoparametric to contour 1.

In terms of the traditional isoperimetric quotient, Theorem 10.1 states that two regions with the same isoperimetric quotient can be scaled to have equal perimeters and equal areas.

We call two contours parametrically similar if they have the same contour ratio. Isoparametric contours are always parametrically similar, whereas parametrically similar contours can be scaled to become isoparametric. Stated another way, isoparametric contours scaled differently are parametrically similar. A pentagram and the special rectangle of base $b \approx 9.6521$ and altitude 1 are parametrically similar but not isoparametric.

For later reference, we find the contour ratios of some specific shapes, with some simple consequences.

Example 1 (Polygons circumscribing a circle). Consider a polygon with perimeter p and area s that circumscribes a circle of radius r . The polygon need

not be regular. Then $s = rp/2$, so the contour ratio of the polygon is

$$\kappa_{\text{poly}} = \frac{p^2}{4s} = \frac{p}{2r}. \tag{10.3}$$

This is the ratio of perimeter to diameter of the inscribed circle (just as π is the ratio of the circumference of a circle to its diameter). For polygons with n sides, the minimum value of κ_{poly} occurs when p is minimal, which means when the polygon is regular. In this case $\kappa = n\tan(\pi/n)$, a bench mark for all n-gons.

The perimeter and area of a circumscribing polygon can, in turn, be calculated in terms of the contour ratio. Using (10.3) we find that

$$p = 2r\kappa_{\text{poly}}, \quad s = r^2\kappa_{\text{poly}}. \tag{10.4}$$

These generalize the classical formulas $p = 2r\pi$ and $s = r^2\pi$ for the circumference and area of the circumscribed circle. From the formulas in (10.4) we conclude:

Polygons with equal contour ratios that circumscribe the same circle are isoparametric. Polygons with equal perimeters that circumscribe the same circle are isoparametric. Polygons with equal areas circumscribing the same circle are isoparametric.

Figure 10.6 shows an equilateral triangle and a rhombus that circumscribe the same unit circle. Each has perimeter $6\sqrt{3}$ and area $3\sqrt{3}$. A circle can be circumscribed by a regular n-gon and by a regular m-gon. If $m \neq n$, the regular polygons necessarily have different perimeters, different areas, and different contour ratios. But an n-gon can be isoparametric to an m-gon if the one with the larger number of sides is not regular, as illustrated by the example in Figure 10.6.

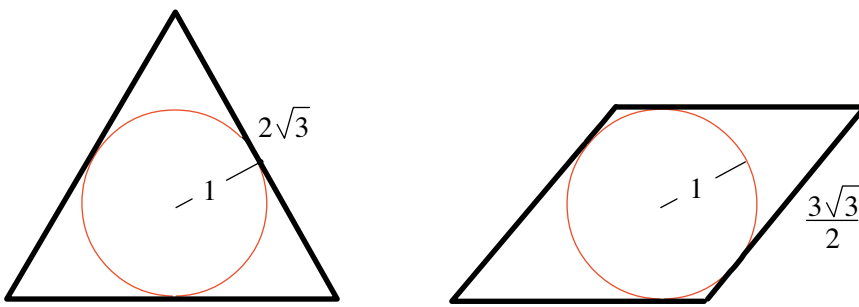


Figure 10.6: An equilateral triangle and an isoparametric rhombus circumscribing the same circle.

Polygons that circumscribe circles have a number of remarkable properties, some of which are alluded to in Example 1. For a more comprehensive list of such properties see Chapter 4.

Example 2 (Rectangle). A rectangle of base b and altitude a has area ab and perimeter $2(a + b)$, so its contour ratio is

$$\kappa_{\text{rect}} = \frac{(a+b)^2}{ab} = 2 + \frac{a}{b} + \frac{b}{a} = 2 + \gamma + \frac{1}{\gamma}, \tag{10.5}$$

where γ is the ratio of the two edges. The minimum occurs when $\gamma = 1$, which gives a square with $\kappa = 4$, a bench mark for all rectangles.

From (10.5) we find that two rectangles with edge ratios γ_1 and γ_2 have the same contour ratio if and only if $\gamma_1 = \gamma_2$ or $\gamma_1\gamma_2 = 1$. In both cases the rectangles are similar. Therefore, dissimilar rectangles cannot be isoparametric.

Example 3 (Circular sector). A circular sector of unit radius subtending an angle of 2θ radians has area θ and perimeter $2 + 2\theta$. Therefore its contour ratio is

$$\kappa_{\text{sect}} = \frac{(2 + 2\theta)^2}{4\theta} = 2 + \theta + \frac{1}{\theta}. \tag{10.6}$$

It is both surprising and remarkable that this has the same form as (10.5), with γ replaced by θ. As illustrated in Figure 10.3a, each sector is isoparametric to a rectangle. Again, the minimum is $\kappa = 4$, the contour ratio of a square, and it occurs when $\theta = 1$. A circular sector subtending an angle of 2 radians appears as a bench mark in Figure 10.5. Sectors subtending angles greater than or less than 2 radians have contour ratio $\kappa > 4$.

Example 4 (Isosceles triangle). An isosceles triangle with equal legs d and vertex angle 2θ has area $s = d^2\sin\,\theta\,\cos\,\theta$ and perimeter $p = 2d + 2d\,\sin\,\theta$, so its contour ratio is

$$\kappa_{\text{isos}} = \frac{4d^2(1 + \sin\,\theta)^2}{4d^2\sin\,\theta\,\cos\,\theta} = \frac{1}{\cos\,\theta}(2 + \sin\theta + \frac{1}{\sin\,\theta}). \tag{10.7}$$

As expected, the right-hand side has its minimum value when $\theta = \pi/6$, giving $\kappa = 3\sqrt{3}$ for an equilateral triangle, a bench mark for all isosceles triangles.

10.3 ISOPARAMETRIC CONTOURS OF DIFFERENT SHAPES

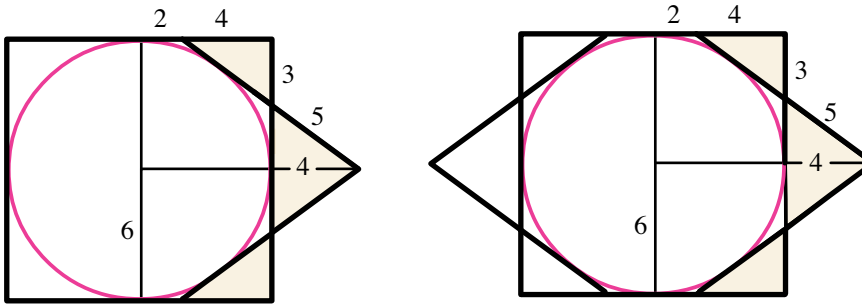
This section solves some special cases of the following type of problem:

Isoparametric contour problem. *Given a contour of specified shape (such as a rectangle), under what conditions can we find an isoparametric contour of another specified shape, such as an isosceles triangle*?

A necessary condition that the shapes be isoparametric is that they have the same contour ratio. If the contour ratios are equal the shapes are parametrically similar, and they can be scaled to become isoparametric.

Example 5 (Square, pentagon, and hexagon). Given a square, whose contour ratio is 4, we wish to find an isoparametric pentagon, which necessarily has contour ratio 4. A general pentagon involves many parameters (such as angles and lengths of edges). We can restrict some of them and still satisfy the requirement that the pentagon should have contour ratio 4. It is easier to find a pentagon that circumscribes the same circle as the given square and has the same perimeter. (A regular pentagon will not do because it has contour ratio smaller than 4.) Figure 10.7a shows an example that works. A circle of radius 6 is circumscribed by a square of side-length 12. Two Pythagorean 3:4:5 triangles are cut off at two corners of the

square by tangents to the circle and flipped to form two edges of a circumscribing pentagon. Because the square and pentagon have equal areas, they are isoparametric. In fact, the circumscribing pentagon has two edges of length 10, two of length 8, and one of length 12, so its perimeter is 48, the same as that of the square. Both the square and pentagon have area 144. If the other two corners of the square are cut off in a similar manner, we find an example of a circumscribing hexagon, shown in Figure 10.7b, that is isoparametric to both the square and pentagon.



(a) Isoparametric square and pentagon (b) Isoparametric square and hexagon

Figure 10.7: Isoparametric square, pentagon, and hexagon circumscribing the same circle.

Example 6 (General contour and rectangle). Assume that we are given a general contour with contour ratio κ, and that we seek an isoparametric rectangle. If $\kappa < 4$ there is no rectangle isoparametric to the given contour because $\kappa \geq 4$ for every rectangle. So for this problem, $\kappa \geq 4$ is a constraint on the general contour. A necessary condition that they be isoparametric is that $\kappa_{\text{rect}} = \kappa$, and from (10.5) we find that $\gamma^2 + (2 - \kappa)\gamma + 1 = 0$, where γ is the side ratio of the rectangle. For $\kappa \geq 4$ the quadratic equation for γ has positive roots given by

$$\gamma = \frac{1}{2}(\kappa - 2) \pm \frac{1}{2}\sqrt{\kappa(\kappa - 4)} = \frac{1}{4}(\sqrt{\kappa} \pm \sqrt{\kappa - 4})^2. \tag{10.8}$$

The product of the roots is 1, so the roots are reciprocals. If $\kappa = 4$, then $\gamma = 1$ and the rectangle is a square. If $\kappa > 4$ there are two distinct roots, γ and $1/\gamma$, but geometrically they are obtained by interchanging the base and altitude of the same rectangle. The rectangle is parametrically similar to the given contour. By Theorem 10.1 the latter can be scaled to become isoparametric to the former.

There is an equivalent way to treat this problem. Suppose that the rectangle has base b and altitude a, and that the given contour has area s and perimeter p. There is no loss of generality if we take $p = 2$, which makes $\kappa = 1/s$. Equating perimeters and areas we find $a + b = 1$ and $ab = a(1 - a) = s = 1/\kappa$, so

$$a(1 - a) = \frac{1}{\kappa}, \tag{10.9}$$

a quadratic in a with two roots, $(1 \pm \sqrt{1 - 4/\kappa})/2$, whose sum is 1. The constraint

$\kappa \geq 4$ ensures that both roots are positive. If we take $a \leq b$, then a is given by

$$a = \frac{1}{2} - \frac{1}{2}\sqrt{1 - \frac{4}{\kappa}} = \frac{1}{2} - \frac{1}{2}\sqrt{1 - 4s}. \tag{10.10}$$

and $b = 1 - a$. We will use this relation in Example 9.

Example 7 (General contour and circular sector). We take a contour with contour ratio κ and try to find an isoparametric circular sector of unit radius. Denote the central angle of the sector, measured in radians, by 2θ. If $\kappa < 4$ there is no sector isoparametric to the given contour because $\kappa_{\text{sect}} \geq 4$ for every sector. Therefore, as in Example 6, $\kappa \geq 4$ is a constraint on the general contour. A necessary condition that the two be isoparametric is that $\kappa_{\text{sect}} = \kappa$, hence by (10.6) we find that $\theta^2 + (2 - \kappa)\theta + 1 = 0$, the same quadratic equation satisfied by γ in Example 6, whose roots are given by (10.8). We know from Example 3 that the sector is isoparametric to a rectangle with edges 1 and θ, so this problem is equivalent to Example 6. For $\kappa \geq 4$, the quadratic has two positive roots that are reciprocals. If $\kappa = 4$, then $\theta = 1$ and there is one circular sector of unit radius subtending an angle of 2 radians. It is isoparametric to the unit square and can be scaled to become isoparametric to any contour with $\kappa = 4$.

If $\kappa > 4$, there are two dissimilar circular sectors of unit radius with the same κ; one subtends an angle of 2θ radians, the other an angle of $2/\theta$ radians. The two sectors are parametrically similar, but (except for the case $\theta = 1$) they are not isoparametric. But if we scale the second sector by a factor θ we obtain a similar sector that is isoparametric to the first sector and also to any contour of contour ratio κ.

In particular, if the contour is a rectangle with edges 1 and θ, there are two sectors isoparametric it. They are shown in Figure 10.3a. Because having contour ratio 4 is a common bench mark for both sectors and rectangles, there are no restrictions on the parameter θ that defines the sector or the rectangle.

Example 8 (General contour and isosceles triangle). Here we start with a contour of contour ratio κ and seek an isoparametric isosceles triangle involving a parameter $t = \sin\theta$, where θ is half the vertex angle. Because $\kappa_{\text{isos}} \geq 3\sqrt{3}$ this problem places the constraint $\kappa \geq 3\sqrt{3}$ on the general contour. A necessary condition that the given contour and an isosceles triangle be isoparametric is that $\kappa_{\text{isos}} = \kappa$. Use (10.7) for κ_{isos} with $\sin\theta$ replaced by t to obtain

$$\kappa = \frac{1}{\sqrt{1 - t^2}} \frac{(1 + t)^2}{t} = \sqrt{\frac{1 + t}{1 - t}} \frac{1 + t}{t}.$$

Now square both sides and rearrange terms to get

$$(1 + \kappa^2)t^3 + (3 - \kappa^2)t^2 + 3t + 1 = 0,$$

a cubic equation for t in terms of κ. When $t = 0$, the cubic polynomial on the left has the value 1, and when $t = -1$, it has the value $-2\kappa^2$, so it always has a root in the interval $(-1, 0)$. A negative root t does not correspond to a possible vertex angle 2θ, so we ignore it.

When $\kappa = 3\sqrt{3}$ the cubic equation becomes $(2t-1)^2(7t+1) = 0$, which has a double root $t = 1/2$ (and a negative root). The root $t = 1/2$ corresponds to $\theta = \pi/6$, which makes the triangle equilateral. When $\kappa > 3\sqrt{3}$, it can be shown that the cubic has two positive roots in the interval (0, 1). For example, when $\kappa = 3 + 2\sqrt{2}$ the equation becomes

$$(t - \frac{1}{2}\sqrt{2})[(18 + 12\sqrt{2})t^2 - (2 + 3\sqrt{2})t - \sqrt{2}] = 0.$$

This cubic has a root at $t = \sqrt{2}/2$, corresponding to $2\theta = \pi/2$, which gives an isosceles right triangle. The quadratic factor has only one positive root $t = 0.3093...$, corresponding to another isosceles triangle with a vertex angle of approximately 36^o.

Example 9 (Rectangle and specified triangular shape). Suppose we want to find an isosceles triangle that is isoparametric to a given rectangle. If the isosceles triangle has base c and equal legs of length d, then $d + c/2 = p/2$ and $s = ch/2$, where h is the altitude of the triangle, given by

$$h^2 = d^2 - (\frac{c}{2})^2 = (d - \frac{c}{2})(d + \frac{c}{2}).$$

There is no loss in generality in assuming that both the rectangle and isosceles triangle have perimeter 2, which means that $\kappa = 1/s$. Then $d + c/2 = 1$ and $d - c/2 = 1 - c$. Hence

$$4s = 2ch = 2c\sqrt{1-c},$$

and (10.10) becomes

$$a = \frac{1}{2} - \frac{1}{2}\sqrt{1 - 2c\sqrt{1-c}}, \tag{10.11}$$

where $0 \le c \le 1$. Figure 10.8 shows the graph of a, plotted as a function of c. The maximal a occurs for $c = 2/3$ (when the triangle is equilateral) and is given by

$$a_{\max} = \frac{1}{2} - \frac{1}{6}\sqrt{9 - 4\sqrt{3}} \ = 0.2601\ldots.$$

This upper bound on a also follows directly from (10.10) by noting that $\kappa \ge 3\sqrt{3}$, the bench mark for all isosceles triangles. It places a constraint on the rectangle parameter a.

For each value of a satisfying $0 < a < a_{\max}$ there are two values of c giving the same a, say $c_1 < c_2$. This means that (with one exception corresponding to $a_{\max}$) there are two different isosceles triangles isoparametric to the rectangle. An example is given in Figure 10.3b, which shows a rectangle with base 6 and altitude 2, together with two isoparametric isosceles triangles. One of them, with base 6, altitude 4, perimeter 16, and area 12, is formed from two Pythagorean $3:4:5$ triangles. When these are scaled by the factor $1/8$ we get an isosceles triangle with perimeter 2 and area $3/16$, so $c = 3/4$ and $a = 1/4$. This gives the point $(3/4, 1/4)$ on the graph in Figure 10.8. The line $a = 1/4$ intersects the graph at a second point $(c_0, 1/4)$, where $c_0 = (1+\sqrt{13})/8 = 0.5757....$ This corresponds to the second isosceles triangle in Figure 10.3b (scaled by a factor 8).

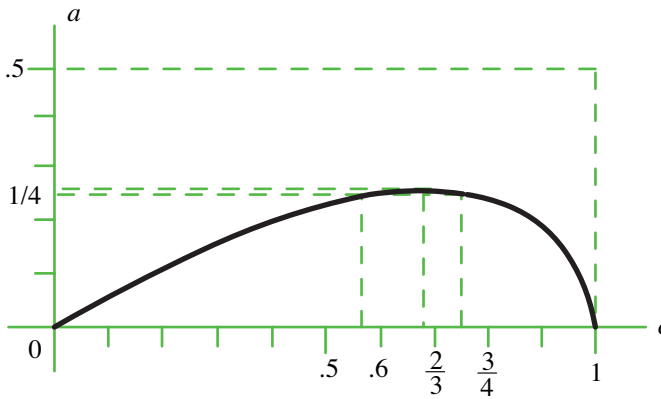


Figure 10.8: Graph of a as a function of c, where a is given by (10.11).

Similarly, suppose that a rectangle with base b and altitude a is given, and that we seek an isoparametric right triangle with base c and altitude h. Again we assume that $a \le b$ and require the perimeter to be 2, so that $a + b = 1$ and $c + h + \sqrt{c^2 + h^2} = 2$. The last equation yields $h = 2(1 - c)/(2 - c)$, and (10.10) becomes

$$a = \frac{1}{2} - \frac{1}{2}\sqrt{1 - \frac{4c(1-c)}{2-c}}. \tag{10.12}$$

The graph of (10.12), with a plotted as a function of c, resembles that in Figure 10.8. The maximal a occurs when $c = h = 2 - \sqrt{2}$ (isosceles right triangle) and is given by

$$a_{\max} = \frac{1}{2} - \frac{1}{2}\sqrt{8\sqrt{2} - 11} = 0.21995\ldots.$$

Again, this is a constraint on the rectangle parameter a. For each a with $0 < a < a_{\max}$ there are two values of c giving rise to the same a; they correspond to two right triangles with their legs interchanged. For small a, the graphs of both (10.11) and (10.12) are almost linear with slope $1/2$. There is a geometric reason for this. When a is very small the rectangle's base b is close to 1 and its area is nearly equal to a. And for small c the triangle's altitude h is close to 1 and its area is nearly $c/2$. Equating areas, we have $a \approx c/2$ for small a and c.

Each of the foregoing examples involves shapes described by a parameter that can be adjusted to make two contour ratios equal. This makes the shapes parametrically similar, so they can be scaled to become isoparametric. We turn next to special contours called rings that also depend on a single parameter.

10.4 RING RATIOS

This chapter treats the simplest type of ring, the region between two similar simple closed curves with similarity ratio λ, where $0 < \lambda < 1$. An example is shown in Figure 10.9. We call λ the size factor because it measures the size of the inner

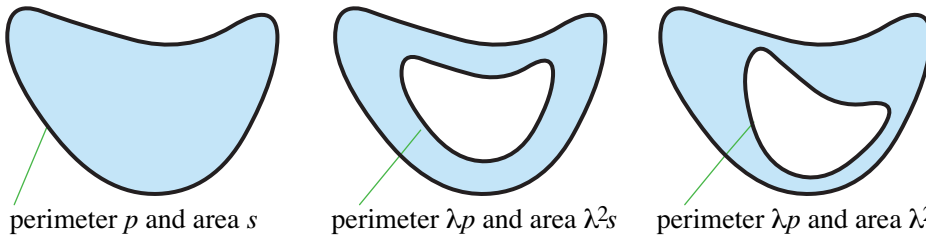


Figure 10.9: A closed curve with perimeter p and area s used to form rings with size factor λ. The inner contour can be located anywhere inside the outer contour.

contour relative to the outer one. If the outer contour has perimeter p and area s, the inner contour has perimeter λp and area $\lambda^2 s$. The inner and outer contours, being similar, have the same contour ratio $\kappa = p^2/4s$.

The ring has its own contour ratio $P^2/(4S)$,where $P = p + \lambda p$ is the total perimeter (the sum of the outer and inner perimeters) and $S = s - \lambda^2 s$ is its area (the difference of the outer and inner areas). The contour ratio of the ring is related to the contour ratio of the inner and outer curves by

$$\frac{P^2}{4S} = \frac{p^2(1+\lambda)^2}{4s(1-\lambda^2)} = \frac{p^2(1+\lambda)}{4s(1-\lambda)} = \kappa\frac{1+\lambda}{1-\lambda}.$$

We call this the *ring ratio* and denote it by ρ to distinguish it from the contour ratio of the boundary curves. Thus we have

$$\rho = \kappa\frac{1+\lambda}{1-\lambda}, \tag{10.13}$$

where κ is the contour ratio for each closed curve forming the ring.

Similar rings have the same ring ratio and the same size factor. Moreover, if two rings are formed from different boundary curves with the same contour ratio κ, then from (10.13) we see that for the same size factor λ they also have the same ring ratio.

Equation (10.13) also shows that

$$\rho > \kappa, \tag{10.14}$$

and, in fact, the ring ratio is always greater than κ by the factor

$$\frac{1+\lambda}{1-\lambda} = 1 + \frac{2\lambda}{1-\lambda}.$$

From (10.13) we find, as well, that λ is uniquely determined by ρ and κ:

$$\lambda = \frac{\rho - \kappa}{\rho + \kappa} = 1 - \frac{2\kappa}{\rho + \kappa}. \tag{10.15}$$

Of course, isoparametric rings have the same ring ratio, but not conversely. In fact, when Theorem 10.1 is applied to rings we obtain:

Corollary 10.1. (Isoparametric Ring Theorem) *Two rings can be scaled to become isoparametric if and only if they have the same ring ratio.*

Figure 10.10 shows various rings arranged as they would appear in Figure 10.5. Rings of different shapes can have the same ring ratio. Rings joined with sloping lines in Figure 10.10 have the same size factor λ. When λ tends to zero, the inner

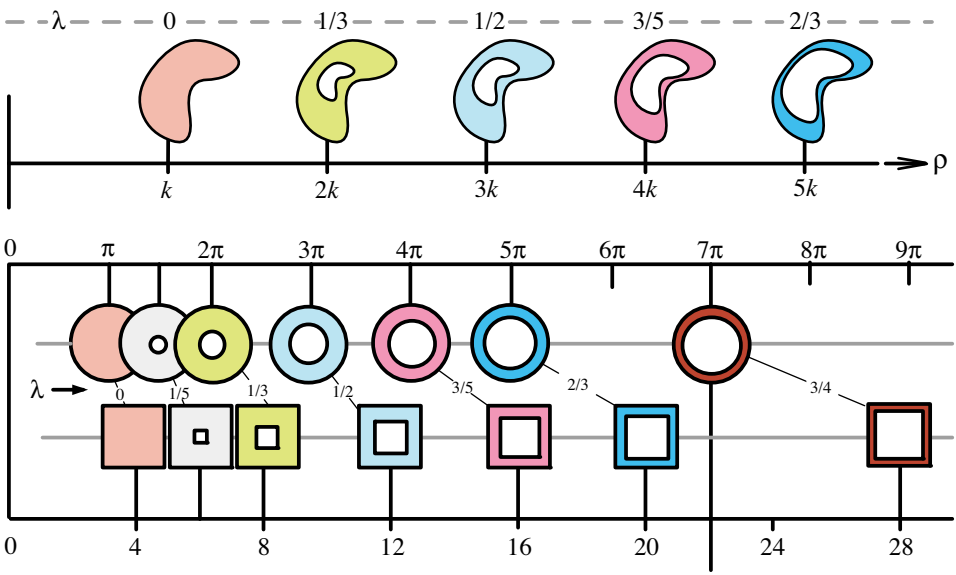


Figure 10.10: Ring ratios of various rings.

contour shrinks to a point, and (10.13) shows that the ring ratio ρ becomes the contour ratio κ, which serves as a bench mark with respect to the parameter λ. In particular, the ring ratio of every circular ring is greater than π and can be made arbitrarily close to π by allowing the hole to shrink to a point. In the same way, the ring ratio of every square ring is greater than 4 and can be made arbitrarily close to 4 by allowing the inner square to shrink to a point. Furthermore, (10.13) implies that $\rho \to \infty$ when $\lambda \to 1$.

10.5 ISOPARAMETRIC INEQUALITY FOR RINGS

Theorem 10.2. *Among all rings with size factor λ, the circular ring has the smallest ring ratio. Hence the ring ratio ρ of a ring with size factor λ satisfies*

$$\rho \geq \pi \frac{1+\lambda}{1-\lambda}, \tag{10.16}$$

with equality only for a circular ring.

Proof. From (10.13) we infer that among all rings with size factor λ the ring ratio is smallest when κ is smallest, and this occurs when $\kappa = \pi$ and the ring is circular. The isoperimetric inequality (10.1) can be regarded as the limiting case of (10.16) when $\lambda \to 0$.

A general ring ratio ρ as defined by (10.13) is a function of κ and λ, which we can denote by $\rho_\kappa(\lambda)$. Inequality (10.16) is a universal inequality,

$$\rho_\kappa(\lambda) \geq \rho_\pi(\lambda),$$

that holds for all rings with size factor λ. There are also local inequalities of the same type for specified rings. For example, $\kappa = 4$ for a square ring, so the ring ratio of a ring with size factor λ and contour ratio $\kappa \geq 4$ satisfies $\rho_\kappa(\lambda) \geq \rho_4(\lambda)$. And, for all rings formed from contours with contour ratio κ we also have the inequality $\rho_\kappa(\lambda) > \kappa = \rho_\kappa(0)$.

Inequality (10.16) tells us when it is possible to have a square ring and circular ring that are isoparametric. We can also see this visually by the display in Figure 10.10. The ring ratio of every square ring exceeds 4, but there are circular rings with ring ratio arbitrarily close to π, so no square ring has ring ratio in the interval between π and 4. For that interval, the hole in the circular ring is too small to make a significant contribution to the total perimeter and area. This puts a constraint on the circular ring in the form of a lower bound on the size factor. If the size factor exceeds this lower bound, then for any square ring there is a circular ring with a large enough hole to match ring ratios.

To describe this quantitatively, suppose that we are given a circular ring and that we ask for an isoparametric square ring. The rings necessarily have the same ring ratio ρ and, moreover, $\rho > 4$. We know from (10.15) that the size factor of a circular ring is given by

$$\lambda = 1 - \frac{2\pi}{\rho + \pi} > 1 - \frac{2\pi}{4 + \pi} = \frac{4 - \pi}{4 + \pi}.$$

In other words, the square ring exists only if the size factor of the circular ring is constrained by the inequality

$$\lambda > \frac{4 - \pi}{4 + \pi} \approx 0.1202\ldots. \tag{10.17}$$

This tells us that the hole in the circular ring should have radius slightly more than 12% of the outer radius. Moreover, if λ satisfies (10.17) then (as will be

demonstrated in Section 10.6) there always exists a square ring isoparametric to the circular ring. The example in Figure 10.3b has $\lambda = 1/2$, which satisfies (10.17). But $\lambda = 1/9$ does not satisfy (10.17).

When we treat the problem in its general form, a constraint of the form

$$\lambda > \frac{\kappa_2 - \kappa_1}{\kappa_2 + \kappa_1} \tag{10.18}$$

will appear, generalizing (10.17). If the rings have equal contour ratios, as do the sectorial ring and rectangular ring in Figure 10.4a, (10.18) becomes $\lambda > 0$, which puts no new constraint on λ. This explains the difference between the examples in Figures 10.4a and 10.4b.

10.6 ISOPARAMETRIC RINGS

We turn next to a general problem motivated by the example of the square ring and circular ring in Section 10.5:

Isoparametric ring problem. *Given a ring with contour ratio κ_1 and size factor λ_1, under what conditions does there exist an isoparametric ring with contour ratio κ_2 and some size factor λ_2?*

The key to this problem is equality of the ring ratios:

$$\rho_1 = \kappa_1 \frac{1 + \lambda_1}{1 - \lambda_1}, \qquad \rho_2 = \kappa_2 \frac{1 + \lambda_2}{1 - \lambda_2}.$$

If $\rho_1 \neq \rho_2$, there is no solution. Therefore, we seek conditions ensuring that

$$\kappa_1 \frac{1 + \lambda_1}{1 - \lambda_1} = \kappa_2 \frac{1 + \lambda_2}{1 - \lambda_2}. \tag{10.19}$$

The problem splits naturally into two cases: $\kappa_1 = \kappa_2$ and $\kappa_1 \neq \kappa_2$.

Case 1: $\kappa_1 = \kappa_2$. In this case, (10.19) holds if and only if $\lambda_1 = \lambda_2$, in which event the rings can be scaled to become isoparametric. In other words, for any two contours with the same contour ratio, say a pentagram and the long narrow rectangle with the same contour ratio mentioned in Section 10.2, we can scale the narrow rectangle to get a similar rectangle isoparametric to the pentagram. This is always possible because of Theorem 10.1. Using these isoparametric contours as outer contours, we scale each of them by the same size factor $\lambda < 1$ to obtain two isoparametric rings. There are infinitely many solutions because we can use any $\lambda < 1$. Therefore, the case $\kappa_1 = \kappa_2$ presents no difficulties and can be regarded as trivial.

Case 2: $\kappa_1 \neq \kappa_2$. We label the contour ratios so that the smaller one is κ_1. Now we have $\kappa_2 > \kappa_1$, and we want to satisfy (10.19). If λ_1 is given, where $0 < \lambda_1 < 1$, and we solve (10.19) for λ_2, then we discover that

$$\lambda_2 = \frac{\kappa_1(1 + \lambda_1) - \kappa_2(1 - \lambda_1)}{\kappa_1(1 + \lambda_1) + \kappa_2(1 - \lambda_1)}. \tag{10.20}$$

But we also need the inequality $0 < \lambda_2 < 1$. The denominator in (10.20) is always positive, but the numerator is positive only if $\kappa_1(1+\lambda_1) > \kappa_2(1-\lambda_1)$. This puts a constraint on λ_1, namely,

$$\lambda_1 > \frac{\kappa_2 - \kappa_1}{\kappa_2 + \kappa_1}. \tag{10.21}$$

Therefore, if λ_1 satisfies the constraint (10.21), then we can always find λ_2 to satisfy (10.19) and we also have $0 < \lambda_2 < 1$. On the other hand, if λ_2 is given and we solve (10.19) for λ_1, then we arrive at a companion result to (10.20),

$$\lambda_1 = \frac{\kappa_2(1+\lambda_2) - \kappa_1(1-\lambda_2)}{\kappa_1(1+\lambda_2) + \kappa_2(1-\lambda_2)}. \tag{10.22}$$

Again, we need $0< \lambda_1 < 1$. The denominator in (10.22) is positive, and the numerator is positive only if $\kappa_2(1 + \lambda_2)> \kappa_1(1-\lambda_2)$, which translates to

$$\lambda_2 > \frac{\kappa_1 - \kappa_2}{\kappa_2 + \kappa_1}.$$

This is automatically satisfied because $\kappa_1 - \kappa_2 < 0$, so there is no constraint on λ_2. Therefore, for given λ_2 we can always find λ_1 to satisfy (10.19), and we get $\rho_1 = \rho_2$. Moreover, from (10.14) we have $\rho_1 > \kappa_2$, or

$$\kappa_1 \frac{1+\lambda_1}{1-\lambda_1} > \kappa_2,$$

which, in turn, is equivalent to (10.21). This means that when we solve for λ_1 using (10.22) it automatically satisfies inequality (10.21). Incidentally, because of (10.15), the relations (10.22) and (10.20) can be expressed more simply in terms of the ring ratio $\rho = \rho_1 = \rho_2$:

$$\lambda_1 = \frac{\rho - \kappa_1}{\rho + \kappa_1}, \quad \lambda_2 = \frac{\rho - \kappa_2}{\rho + \kappa_2}.$$

The foregoing results are summarized in the following theorem.

Theorem 10.3. *Two rings with contour ratios κ_1 and κ_2 ($\kappa_2 \geq \kappa_1$) and size factors λ_1 and λ_2 can have the same ring ratio ρ if and only if*

$$\lambda_1 > \frac{\kappa_2 - \kappa_1}{\kappa_2 + \kappa_1}. \tag{10.23}$$

In this case, the size factor λ_2 is uniquely determined by the equation

$$\lambda_2 = \frac{\rho - \kappa_2}{\rho + \kappa_2}. \tag{10.24}$$

When $\kappa_2 = \kappa_1$, constraint (10.23) is automatically satisfied, and (10.24) gives $\lambda_2 = \lambda_1$. This case is illustrated by the two sectorial rings and the rectangular ring shown in Figure 10.4a. All three rings have contour ratio 4 and equal size factor λ. Infinitely many pairs of two sectorial rings or of a sectorial ring and a rectangular ring are obtained by allowing λ to vary between 0 and 1. Another such example

comes from the pentagram and the special rectangle mentioned in the foregoing proof of Case 1.

As already noted, when the size factor $\lambda \to 0$ the hole in a ring shrinks to a point. The restriction $\lambda > 0$ is imposed on the size factor in order to produce an inner closed curve similar to the outer one. But formula (10.13) that defines the ring ratio ρ is meaningful if $\lambda = 0$ and gives $\rho = \kappa$ in that case. Consequently, the case $\kappa_1 < \kappa_2$ of Theorem 10.3 is applicable for the limiting value $\lambda_2 = 0$, in which event the constraint on λ_1 in (10.23) becomes an equality. And conversely, if $\lambda_1 = (\kappa_2 - \kappa_1)/(\kappa_2 + \kappa_1)$, then the corresponding value of λ_2 is 0. Thus, for example, if $\kappa_2 = 4$ and $\kappa_1 = \pi$, a circular ring with ring ratio 4 and size factor $\lambda_1 = (4 - \pi)/(4 + \pi)$ can be scaled to become isoparametric to a square, a circular sector, or any other contour with contour ratio 4, each of which can be regarded as a limiting case of a ring with $\lambda_2 = 0$. The reader can verify that the following examples, which illustrate Theorem 10.3, also cover the corresponding limiting cases with $\lambda_2 = 0$ if the constraint inequality in (10.23) is changed to an equality.

Example 10 (Circular ring and regular polygonal ring). The problem for circular rings and square rings considered in Section 10.5 can be generalized by replacing the square with a regular n-gon. Taking $\kappa_0 = \pi$ and $\kappa_n = n \tan(\pi/n)$, we have $\kappa_n > \kappa_0$, so if a circular ring has size factor λ_0 satisfying

$$\lambda_0 > \frac{n \tan \frac{\pi}{n} - \pi}{n \tan \frac{\pi}{n} + \pi},$$

there is always an isoparametric regular n-gon ring. When $n = 4$ this is inequality (10.17). As n increases, the lower bound on λ_0 decreases. For example, when $n = 3$, the lower bound is about 0.2464, but when $n = 12$ it is about 0.0116. This indicates that the radius of the hole in the circular ring needs to be more than 24% of the outer radius to have an isoparametric triangular ring, but 1.2% suffices for a dodecagonal ring. There is more latitude in finding isoparametric dodecagonal rings because a dodecagon is more circular than a triangle.

Example 11 (Pythagorean 3:4:5 triangular ring and square ring). Consider a triangular ring bounded by two similar Pythagorean $3 : 4 : 5$ triangles. An example with outer triangle of edges 9, 12, and 15 and size factor 1/3 is depicted in Figure 10.11. All such rings have contour ratio $\kappa = 6$ and ring ratio $\rho > 6$. An isoparametric square ring with inner side-length a and outer side-length b has contour ratio 4 and the same ring ratio ρ. Since $4 < 6$, we label the square ring as ring 1 and the triangular ring as ring 2. Constraint (10.23) states that isoparametric square rings exist if they have size factor

$$\lambda_2 > (6 - 4)/(6 + 4) = 1/5.$$

The square ring in Figure 10.11 has $\lambda_2 = 1/2$, which suffices. The triangular ring has perimeter 48, so equality of total perimeters dictates that $48 = 12a$, and we find that $a = 4$, $b = 8$. But no square ring with size factor smaller than 1/5 is isoparametric to any 3:4:5 triangular ring.

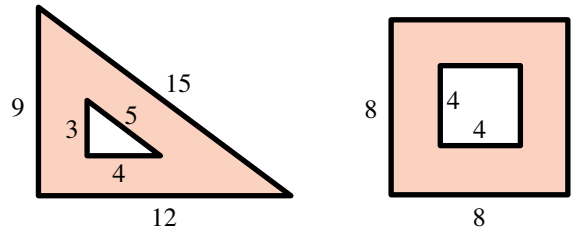


Figure 10.11: A Pythagorean 3:4:5 triangular ring and an isoparametric square ring.

Example 12 (Two regular polygonal rings). This example compares a ring formed by two regular n-gons with one formed by two regular m-gons, where $n > m$. The contour ratio for the n-gon is $\kappa_n = n\tan(\pi/n)$, while that for the m-gon is $\kappa_m = m\,\tan(\pi/m)$. Because κ_n is a decreasing function of n, we have $\kappa_n < \kappa_m$, and (10.23) becomes

$$\lambda_n > \frac{m\,\tan(\pi/m) - n\tan(\pi/n)}{m\tan(\pi/m + \,n\tan(\pi/n)}, \tag{10.25}$$

where λ_n and λ_m replace λ_1 and λ_2 in Theorem 10.3. From the polynomial approximation $\tan\, x \sim x + x^3/3$, valid for small x, we see that for large m and n the quotient on the right of (10.25) has the asymptotic value

$$\frac{\pi^2}{6}(\frac{1}{m^2} - \frac{1}{n^2}).$$

10.7 ISOPARAMETRIC RINGS WITH EQUAL INNER PERIMETERS AND EQUAL OUTER PERIMETERS

Two rings that are isoparametric have the same area and the same total perimeter. We can also construct such rings in which both inner perimeters are equal and both outer perimeters are equal. For example, take two different polygons that circumscribe the same circle and have the same perimeter. Then they are isoparametric and have the same contour ratio. Now scale each polygon by the same size factor λ to produce two polygonal rings that are isoparametric. These rings have the additional property that both the inner perimeters are equal and the outer perimeters are equal. In fact, for each λ we obtain a family of such isoparametric polygonal rings by moving the inner contour, and by varying λ we obtain even more families. This will be used in Section 10.8 to generate a remarkable family of incongruent solids satisfying the properties (a) through (f) listed in Section 10.1. More generally, take two incongruent isoparametric contours bounded by simple closed curves. Scale each of them by the same size factor λ to produce two incongruent isoparametric rings. Then these rings also have the property that both the inner perimeters are equal and the outer perimeters are equal. Figure 10.4a shows such an example. We leave it to the reader to verify that a necessary and sufficient condition for two

isoparametric rings to have both the inner perimeters equal and the outer perimeters equal is that both the two outer contours be isoparametric and the two inner contours be isoparametric.

10.8 INCONGRUENT SOLIDS WITH PROPERTIES (a) TO (f)

This section describes several families of incongruent solids having properties (a) through (f) listed in the introduction. Each family is generated by isoparametric circumscribing polygons of the type discussed in Section 10.7.

Start with a smooth solid of revolution whose cross sections by horizontal planes perpendicular to the rotation axis are circular rings. Take any polygonal ring of the type discussed in Section 10.7 that circumscribes the base, and use a similar polygonal ring to circumscribe each parallel circular cross section above the base. The union of these polygonal rings sweeps out a solid, an example of which is shown in Figure 10.12a.

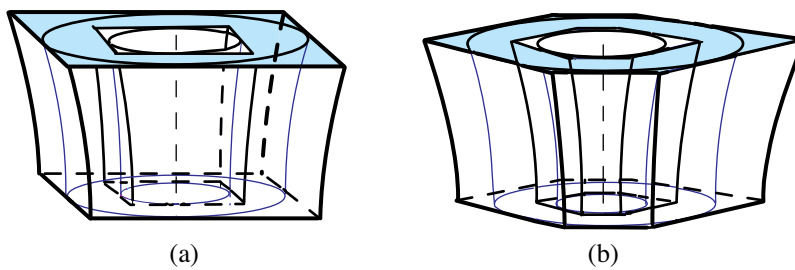


Figure 10.12: Incongruent solids sharing properties (a) through (f) of Section 10.1.

Repeat the process, starting with a noncongruent isoparametric polygonal ring on the base to produce a noncongruent solid like the example in Figure 10.12b. In each horizontal cross section the two polygonal rings are isoparametric and circumscribe the same circle. Moreover, the two inner polygons have equal perimeters, as do the two outer polygons. Equality of cross-sectional areas implies equality of volumes of the two solids, and, because the perimeters are equal, it is not difficult to prove that both the inner and outer lateral surface areas are equal. In each solid, the inner lateral surface is similar to the outer lateral surface by some size factor λ.

We can choose the polygonal rings in infinitely many ways, so for a solid of revolution we have infinitely many pairs of incongruent solids satisfying properties (a) through (f). And we can generate more such families by starting with different solids of revolution. In particular, when the solid of revolution is a hemisphere, the solids are related to the Archimedean shells discussed in Chapter 5. In this case, two different Archimedean shells with isoparametric polygonal bases circumscribing congruent equators of two hemispheres provide examples of noncongruent solids sharing properties (a) through (f). Moreover, the inner solid in Figure 10.12 can be moved as suggested by Figure 10.9 to produce even more examples.

PART 2: DISSECTIONS OF ISOPARAMETRIC REGIONS

10.9 DISSECTIONS INVOLVING BOUNDARIES

It is well known that any planar polygonal region can be dissected into smaller polygonal pieces that can be rearranged to form any other polygonal region of equal area (see for example [41; p. 221]).

Figure 10.13: Dissection (a) of a triangle onto a rectangle; (b) of a rectangle onto another rectangle of prescribed altitude. Dark lines show how the boundaries are transformed.

What happens to the boundaries in these standard dissections?

Figure 10.13a shows a dissection of a triangle onto a rectangle, and Figure 10.13b shows a dissection of one rectangle onto another rectangle of prescribed altitude. In both examples, part of the boundary of one polygon ends up inside the other. This is to be expected, because the initial and final shapes have different perimeters. Figure 10.34a shows an even more dramatic example, a famous hinged dissection of Dudeney in which the entire boundary of a square ends up inside a triangle.

We are interested in dissections that not only preserve areas but also convert the boundaries onto each other as well. We call these *complete dissections*. This requires that both regions have equal areas and equal perimeters, so they are isoparametric. To the best of our knowledge, such dissections have not been previously treated.

In Figure 10.14, the triangle and rectangle are isoparametric, but the dissection shown is not complete because one boundary is not converted entirely onto the other.

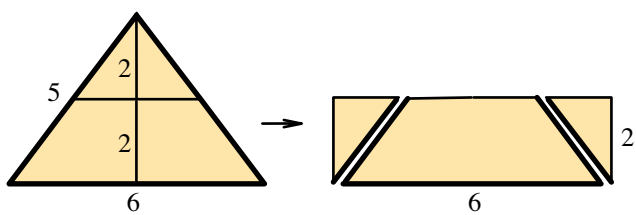


Figure 10.14: Standard dissection converting a triangle onto an isoparametric rectangle.

In fact, there is no reason to expect that a complete dissection exists even though the figures are isoparametric. Nevertheless, our next theorem reveals a surprising and profound result: for any two isoparametric convex polygonal regions, a complete dissection always exists. Moreover, the proof shows how to construct one.

Before proceeding, the reader might try to find a complete dissection that converts the triangle in Figure 10.15 onto the isoparametric isosceles trapezoid shown. This is a simpler task than for the pair in Figure 10.14.

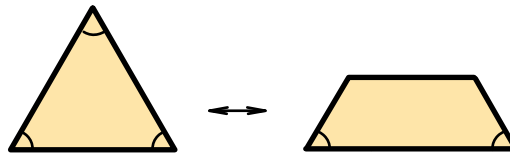


Figure 10.15: Equilateral triangle and an isoparametric trapezoid.

Figure 10.16 shows a complete dissection of an isosceles triangle onto an isosceles trapezoid, and of a general triangle onto a trapezoid with the same base angles.

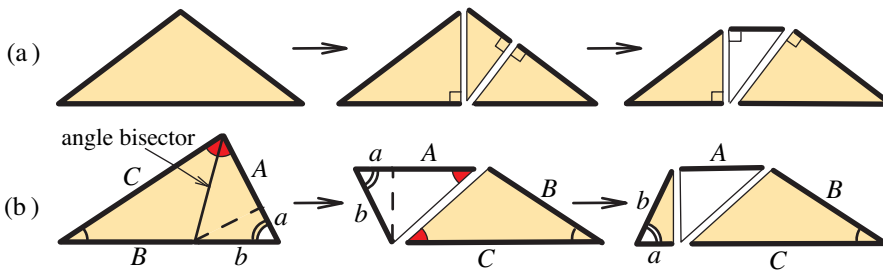


Figure 10.16: Complete dissection converting (a) isosceles triangle onto isosceles trapezoid, and (b) any triangle onto a trapezoid. In (a), the unshaded piece has been flipped. In (b) one piece is flipped and divided into two right triangles, the smaller of which is flipped again.

10.10 COMPLETE DISSECTION OF POLYGONAL REGIONS

We turn now to the first principal result concerning complete dissection.

Theorem 10.4. *Any two isoparametric convex polygonal regions can be converted onto one another by complete dissection.*

Proof. Consider two isoparametric convex polygonal regions A and B. Figure 10.17 shows an example, with A triangular and B quadrilateral; it displays all the essential features required in treating general convex isoparametric polygonal regions. The method of proof is suggested by an oversimplified intuitive idea: Remove each boundary and perform a standard dissection of the interior of A to produce the interior of B. Then restore the two boundaries to obtain the complete dissection.

To make this intuitive idea rigorous, refer to Figure 10.18a, which shows a frame of constant width w protruding into each region along its boundary. Choose w small enough so that the inner boundary of each frame will be a simple closed polygon with the same number of sides. Each region now consists of two parts: the frame plus the interior region it surrounds. Keep in mind that:

The sum of areas, frame plus interior, is the same for both regions A and B.

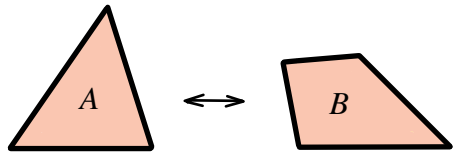


Figure 10.17: Isoparametric triangular and quadrilateral regions.

Now unfold each frame at the outer vertices (thought of as hinges) and lay it out horizontally, as in Figure 10.18b. To be specific, use angle bisectors to cut each frame into trapezoidal pieces with isosceles triangular gaps between adjacent pieces.

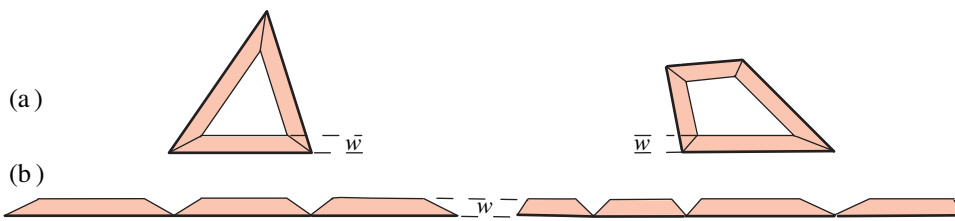


Figure 10.18: Frames unfolded to form adjacent trapezoidal pieces with triangular gaps.

Next, use standard dissections (as in Figure 10.13) of the regions interior to the frames in Figure 10.18a to convert them onto two rectangles with common altitude w (the frame width), as indicated in Figure 10.19a. Each rectangle has area equal to that of the interior that produced it, but, of course, the two rectangular areas are not equal to each other.

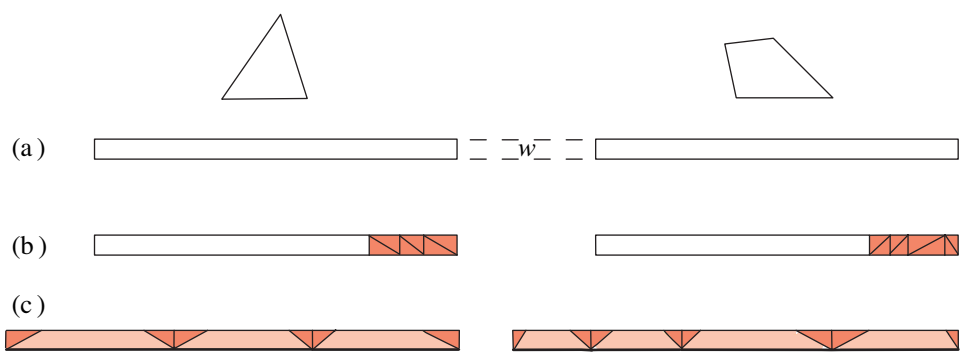


Figure 10.19: (a) Dissection of interior regions produces two rectangles of equal altitude, but not of equal area. Removing shaded triangular pieces in (b) to fill the gaps as shown in (c) leaves two unshaded congruent rectangles in (b).

From each of these rectangles, remove smaller triangles as needed (see Figure 10.19b) to fill the triangular gaps in Figure 10.18b. This transforms the unfolded

frames into two congruent rectangles shown in Figure 10.19c, whose lower bases are the unfolded boundaries of A and B. By overlapping the rectangles we obtain their common dissection (dissection 1) which also converts the full boundary of A onto the full boundary of B. The leftover unshaded rectangles in Figure 10.19b are also congruent because the sum of areas, frame plus interior, is the same for both regions in Figure 10.18a. By overlapping the unshaded rectangles in Figure 10.19b, we obtain a common dissection (dissection 2) of the two interior dissections inherited from Figure 10.19a. The union of common dissections 1 and 2 gives a complete dissection of region A onto region B and completes the proof.

In the foregoing dissections, some pieces may have been flipped. However, as will be demonstrated in Theorem 10.6, all complete dissections can also be done without flipping, although this may require more pieces.

10.11 COMPLETE DISSECTION OF POLYGONAL FRAMES

The frames of constant width used in the proof of Theorem 10.4 lead us to consider the problem of complete dissection of such frames, where now both inner and outer boundaries are subject to conversion. Although frames are more complicated objects than those in Theorem 10.4, our next principal result (Theorem 10.5) states that any two isoparametric frames can also be converted onto one another by complete dissection. First we explain precisely what we mean by a polygonal frame.

In this chapter, the term *polygonal frame* refers to a frame of constant width. It has parallel inner and outer boundaries with constant distance separating the parallel edges. We restrict our discussion to *convex* polygonal frames, that is, frames in which both the inner and outer polygons are convex.

More precisely, start with a convex n-gon as inner boundary, and any frame-width w. Draw lines outside the n-gon parallel to its sides at distance w from the sides. Segments of the lines will form another convex n-gon outside the inner boundary, like those in Figure 10.18a. The frame consists of the region between the two n-gons, including both boundaries. When unfolded at the outer vertices, the frame forms a set of adjacent trapezoidal regions akin to those in Figure 10.18b.

Complete dissections require isoparametric frames, which have equal areas and equal total perimeters (inner plus outer). Figure 10.20 shows an example mentioned in Figure 10.11: a Pythagorean 3:4:5 triangular frame of constant width $w = 2$, and a square frame of the same constant width.

Note that both the areas and total perimeters of these frames have the same numerical value, 48. This is not merely a coincidence, but is a consequence of the following lemma when $w = 2$.

Lemma 10.1. *The width w, area A, and total perimeter P of a convex polygonal frame are related by*

$$A = \frac{1}{2}Pw. \tag{10.26}$$

The proof follows easily by applying the area formula for adjacent trapezoids forming the frame.

As an immediate consequence of Lemma 10.1 we have the following crucial result:

Corollary 10.2. *All isoparametric convex polygonal frames have the same width.*

Now return to Figure 10.20, which shows an unfolding of each frame into trapezoids of constant altitude $w = 2$, followed by a dissection onto a rectangle of the same altitude with two horizontal bases, the sum of whose lengths is the total perimeter of the frame. Because the two frames are isoparametric, so are the two rectangles. At the bottom of Figure 10.20, the dissected rectangles are superim-

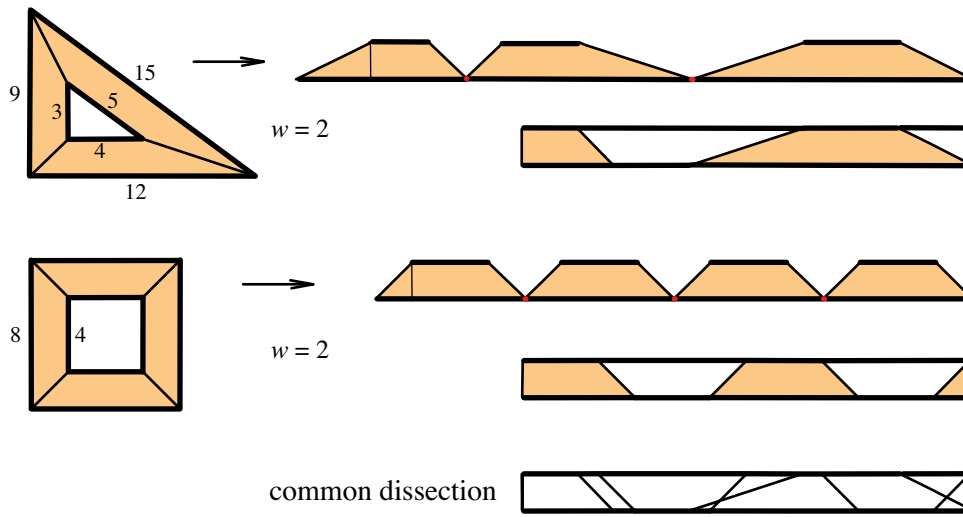


Figure 10.20: Complete dissections of two special isoparametric frames.

posed to obtain a common dissection of the two frames that converts their total boundaries, indicated by the heavy lines.

The same type of argument can be applied to any pair of isoparametric convex polygonal frames to prove the following theorem.

Theorem 10.5. *Any two convex isoparametric polygonal frames can be converted onto one another by complete dissection.*

10.12 COMPLETE DISSECTIONS WITHOUT FLIPPING

In the foregoing dissections, some pieces may have been flipped. Although flipping might reduce the number of dissection pieces, in some applications, such as skin grafting or laying out carpeting, flipping is undesirable. It is known that every standard dissection can be carried out without flipping. Now we show that the same is true for every complete dissection.

Theorem 10.6. *Every complete polygonal dissection can be done without flipping.*

Flipping a piece turns it into its mirror image, so to prove Theorem 10.6 it suffices to show that a complete dissection of a general polygonal piece onto its

mirror image can be done without flipping and in such a way that each edge of the polygon is converted to the corresponding edge of the mirror image. Such a dissection requires more than completeness, and we call it *strongly* complete. Now we will prove that every polygon can be converted onto its mirror image by using strongly complete dissection. First dissect the polygonal piece into triangles, as illustrated in Figure 10.21. Then perform a strongly complete dissection on each triangle. To do this, we can dissect each triangle into six right triangles as shown

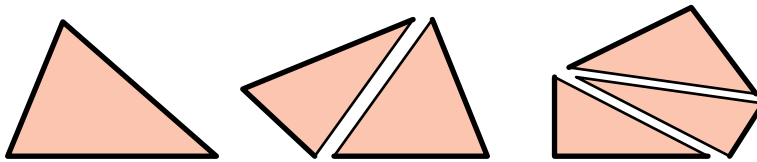


Figure 10.21: Quadrilateral and pentagon cut into triangular pieces.

in Figure 10.22a, where now each right triangle has only one leg as part of the boundary of the original triangle. This reduces the problem to that of dissecting a right triangle onto its mirror image, without flipping, so that one leg gets converted onto its mirror image. Figure 10.22b shows how a right triangle can be cut into two isosceles triangles. Rotate one of them to convert the given leg as shown.

By combining Theorems 10.4 and 10.6 we conclude that two isoparametric convex polygons can be converted onto one another without flipping.

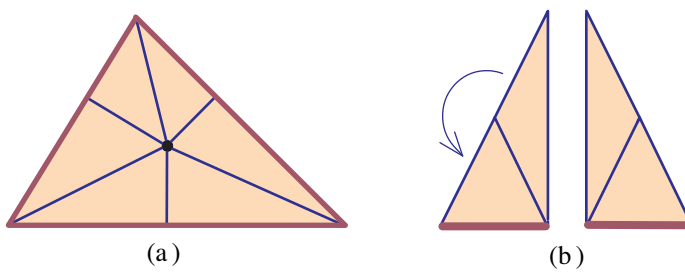


Figure 10.22: (a) Cutting a general triangle into right triangles, each of which has only one leg subject to conversion. (b) Converting one leg of a right triangle onto its mirror image.

Examples. Figure 10.23 shows two different complete dissections converting two isoparametric curvilinear regions onto one another. In the left figure (taken from Figure 10.2), a chord bisects the oval, and one piece is flipped as in the examples of Figure 10.2 to produce the symmetric heart-shaped figure. A complete dissection without flipping is shown on the right, where cuts are made along the diagonals of the square inscribed in the oval, and then the pieces are rotated as shown. This dissection works for any oval having two perpendicular axes of symmetry.

Figure 10.23: Two complete dissections of isoparametric curvilinear regions.

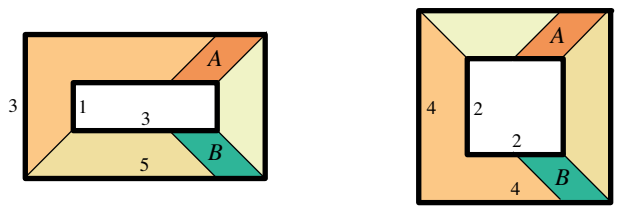


Figure 10.24: Complete dissection of rectangular frame and isoparametric square frame.

Figure 10.24 shows a complete dissection without flipping of a rectangular frame onto an isoparametric square frame. A similar dissection works for any two rectangular frames of equal width having equal outer perimeters.

10.13 COMPLETE DISSECTIONS USED TO APPROXIMATE CURVILINEAR REGIONS

Figure 10.25 shows a frame that is partially polygonal and partially curvilinear. This example consists of a quadrilateral portion together with a curvilinear portion obtained as a limit of a portion of a polygonal frame of the same constant width w. The second region in Figure 10.25 is a circumgon of inradius w that can be

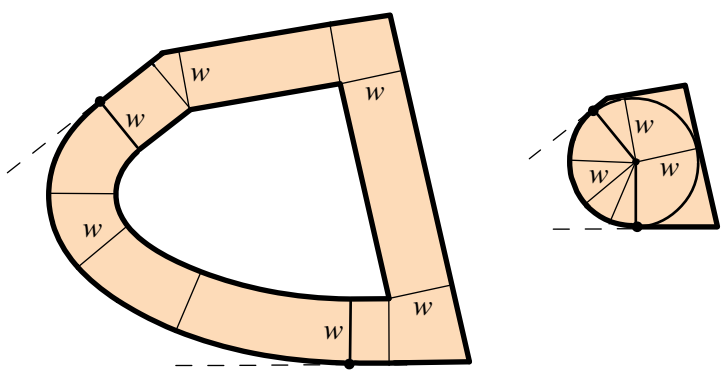


Figure 10.25: Frame with partially curvilinear boundaries, and a circumgon.

regarded as a degenerate case of the frame on the left, with the inner boundary replaced by a single point and the curved portion being a circular arc. Circumgons

were introduced in Chapter 4, where it was shown that the area A and perimeter P of any circumgonal region with inradius w are related by $A = Pw/2$, which is (10.26) in Lemma 10.1. More general frames, partially curvilinear and partially

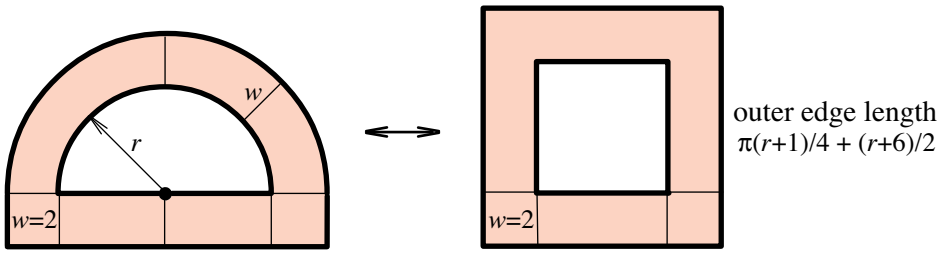


Figure 10.26: A partially semicircular frame and an isoparametric square frame.

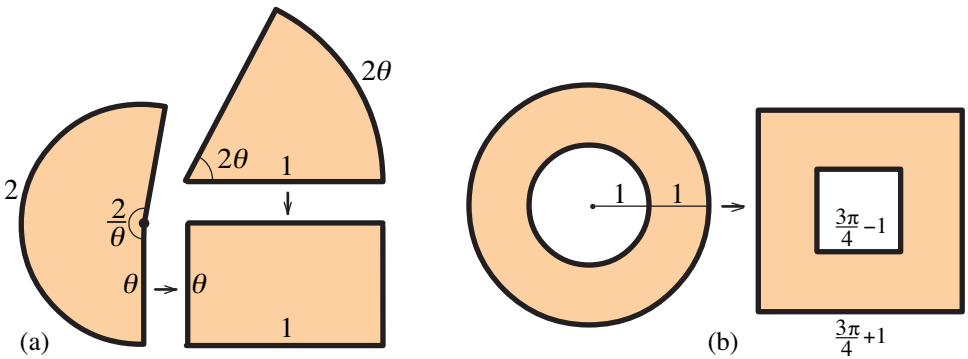


Figure 10.27: Partially circular regions isoparametric to polygonal regions.

polygonal, can be introduced similarly, so that they share formula (10.26) of Lemma 10.1. By Corollary 10.2 any two isoparametric partially curvilinear frames have the same width.

Figure 10.26 shows an example in which the curvilinear part is semicircular and the polygonal part is rectangular. The partially semicircular frame and the square frame shown adjacent to it are isoparametric. Actually, the square frame is only one of infinitely many rectangular frames of width $w = 2$ isoparametric to the semicircular frame. If the inner radius of the semicircular part is r, the common value of the area and the perimeter is $A = P = 2\pi(r + 1) + 4(r + 2)$.

Figure 10.27a (borrowed from Figure 10.3a) shows two isoparametric circular sectors that are also isoparametric to the same rectangle. Figure 10.27b (from Figure 10.4b) shows a circular frame and an isoparametric square frame of the same width.

Figure 10.28 shows how the circular sectors in Figure 10.27a can be dissected into an even number of radial slices that can be rearranged to form a figure approximating the same rectangle. In one case the slices are arranged in horizontal layers,

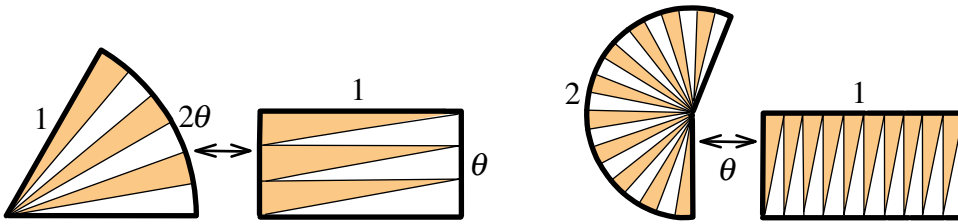


Figure 10.28: Radial slicing of two special circular sectors with their boundaries, rearranged differently to approximate the same isoparametric rectangle.

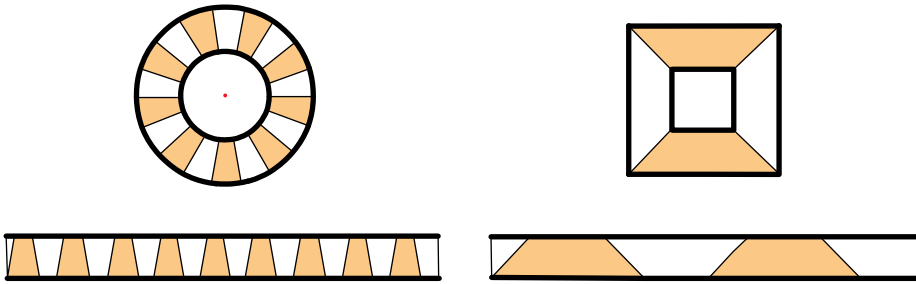


Figure 10.29: Radial slices of a circular frame and its boundary, rearranged to approximate the isoparametric square frame.

and in the other case in vertical layers. As the number of slices increases without bound, both approximations have the same rectangle as a limit, with the curved boundaries becoming the rectangular boundaries, in the style of Archimedes.

Figure 10.29 shows a corresponding dissection of the circular frame in Figure 10.27b that approximates the square frame.

The two frames in Figure 10.27b are members of an infinite family of isoparametric regular frames of width $w = 1$, with examples shown in Figure 10.30.

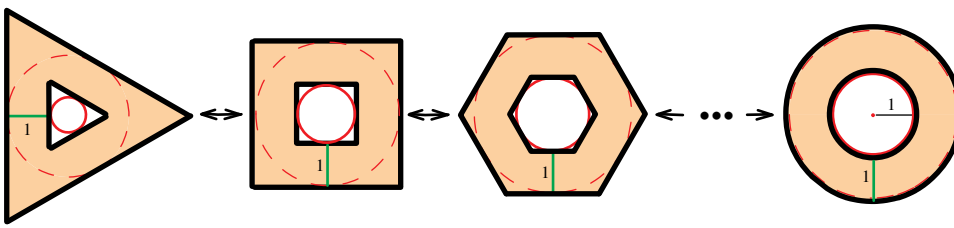


Figure 10.30: Regular polygonal frames and a circular frame, all isoparametric.

Each inner regular n-gon in this family has an incircle of diameter d_n, where

$$d_n = \frac{3\pi}{n \tan \frac{\pi}{n}} - 1,$$

and each outer regular n-gon has an incircle of diameter $D_n = d_n + 2w = d_n + 2$.

10.14 ISOPERIMETRIC PROPERTIES OF FRAMES

This section introduces new isoperimetric properties of convex polygonal frames. First, we prove the following:

Theorem 10.7. *Among all convex isoparametric n-gonal frames, the equiangular n-gonal frames have the properties:*
(a) *the inner polygon has the largest perimeter;*
(b) *the outer polygon has the smallest perimeter.*

Proof. By Corollary 10.2, all isoparametric n-gonal frames have the same width, which we denote by w. We illustrate the proof with a quadrilateral frame, shown in Figure 10.31a. A convex polygonal frame with outer perimeter a and inner perimeter b has total perimeter $P = a + b$. Let $\Delta = a - b$ denote the difference of the outer and inner perimeters. In Figure 10.31a, right-angle cuts at each vertex of the inner boundary divide the frame into four rectangles and four corner pieces, and Δ is the sum of the lengths of the heavy line segments meeting at the four outside vertices.

The four corner pieces can be translated so their inner vertices are brought to a common point that serves as the center of a circle of radius w, as in Figure 10.31b.

The sum of the vertex angles at the inner vertices of the corner pieces is exactly 360°, hence in their new positions, the union of the four corner pieces is a circumgon that fills a quadrilateral region circumscribing a circle of radius w. For a general n-gonal frame, this will be an n-sided circumgon with inradius w.

The perimeter of this circumgon is Δ, and among all n-gonal frames, Δ will have its smallest value when the circumgon is equiangular. This is illustrated for a quadrilateral by comparing perimeters in Figures 10.31b and 10.31c.

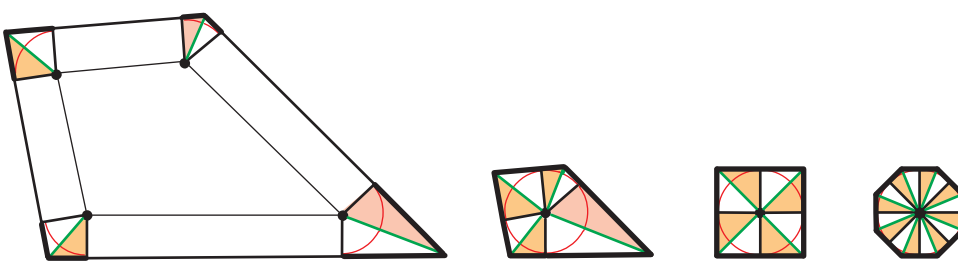


Figure 10.31: The difference Δ between outer and inner perimeters in (a) is the perimeter of a circumgon in (b) with inradius w. This perimeter is smallest when the circumgon is equiangular (a regular polygon), as shown in (c) for $n = 4$, and in (d) for $n = 8$.

The frame in Figure 10.31a has perimeter $P = a+b = 2b+\Delta$, hence $2b = P-\Delta$. So, for a given total frame perimeter P and width w, the inner perimeter b will be largest when Δ is smallest, and this occurs for equiangular n-gonal frames. This

proves Theorem 10.7a. The proof of Theorem 10.7b follows in analogous fashion using the relation $2a = P + \Delta$.

Corollary 10.3. *For given total perimeter P, all equiangular isoparametric frames have the same Δ, hence equal inner perimeters b and equal outer perimeters a.*

By a limiting argument, Theorem 10.7 implies: *Among all isoparametric frames, the circular frame has the largest inner perimeter and the smallest outer perimeter.* Figure 10.30 compares a circular frame with regular polygonal isoparametric frames.

Parallel frames.

Equiangular frames are special cases of *parallel frames*, illustrated in Figure 10.32. Two n-gonal frames are called *parallel* if corresponding edges are parallel, or, equivalently, if corresponding angles of the outer polygons are equal. Each frame has its own constant width, but two parallel frames can have different widths.

Figure 10.32 shows three parallel pentagonal frames of different widths, one of them being a circumgon, which can be regarded as a frame whose inner boundary is a point. Circumgons play a fundamental role among parallel frames, as revealed by the following isoperimetric property.

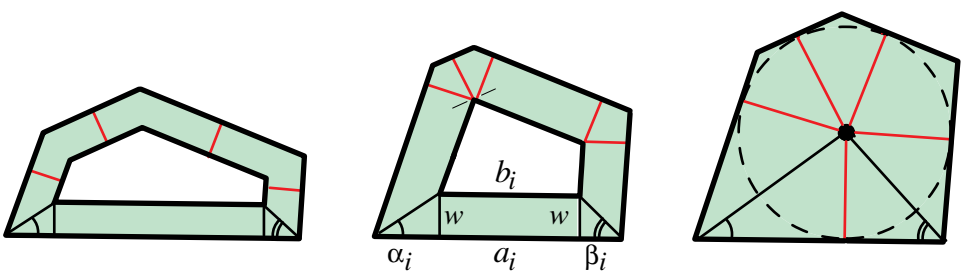


Figure 10.32: Three parallel pentagonal frames of different widths. Corresponding angles at the outer polygons are equal. Among all parallel n-gonal frames of given total perimeter, the circumgon has largest width and therefore largest area.

Theorem 10.8. *Of all parallel convex n-gonal frames with a given total perimeter, the circumgon has the maximal area.*

Proof. For a given convex n-gonal frame, denote the lengths of the outer edges by $a_1, ..., a_n$, and the corresponding lengths of inner edges by $b_1, ..., b_n$, and let $\Delta_i = a_i - b_i$. The total perimeter of the frame is

$$P = \sum_{i=1}^{n}(a_i + b_i) = \sum_{i=1}^{n}(2b_i + \Delta_i),$$

and we are comparing all parallel frames with different widths w for a given value of P. The parallel edges of lengths a_i and b_i form part of a trapezoid of altitude

w and base angles α_i and β_i indicated, respectively, by single arcs and double arcs in Figure 10.32. The difference Δ_i is the sum of the lengths of two segments, $\Delta_i = w\cot\alpha_i + w\cot\beta_i = wc_i$, where $c_i = \cot\alpha_i + \cot\beta_i$. For a given i, c_i is the same constant for all n-gonal parallel frames because corresponding angles of such frames are equal. Therefore

$$P = 2\sum_{i=1}^{n} b_i + w\sum_{i=1}^{n} c_i = 2B + wC,$$

where $B = \sum_{i=1}^{n} b_i$ is the perimeter of the inner polygon, and where $C = \sum_{i=1}^{n} c_i$ has the same value for all parallel frames. Hence

$$w = \frac{P - 2B}{C},$$

from which it follows that for fixed P, w attains its largest value when $B = 0$, that is, when the inner boundary shrinks to a point. This occurs for the circumgon with inradius equal to this maximal width w. By Lemma 10.1, the area of each frame of given perimeter is proportional to its width, so the maximal area occurs for the circumgon, as asserted.

We have verified that Theorem 10.8 can be extended to frames that are partially curvilinear, in which case the curved part of the circumgon is an arc of the incircle.

10.15 DESIGNATED COMPLETE DISSECTIONS

Our proof of Theorem 10.4, which shows how two isoparametric polygonal regions can be converted onto one another by complete dissection, can be modified slightly to prove many surprisingly stronger results. For example, consider two isoparametric convex polygonal regions, each of whose boundaries is divided into the same number of parts of prescribed but arbitrary lengths. Imagine that corresponding parts of equal length are assigned the same color, say r (red), g (green), b (blue), y (yellow), v (violet), as in the examples of Figure 10.33. Then there is a complete

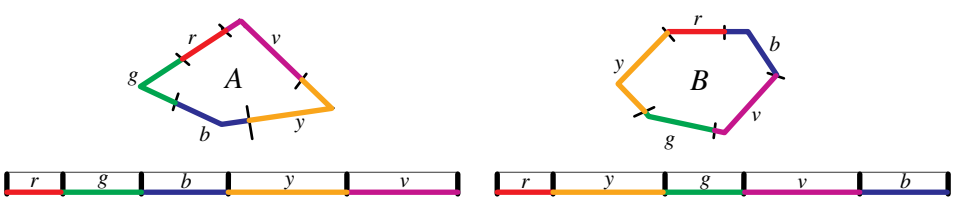


Figure 10.33: Two isoparametric polygons with boundaries divided into parts of prescribed length identified by colors r, g, b, y, v. The unfolded frames with gaps filled form two congruent rectangles with pieces of the same color being congruent.

dissection of the polygons such that all parts of the boundaries are converted onto another as designated by the colors. To see why, unfold the frames and fill the gaps as in the proof of Theorem 10.4 to form two congruent rectangles as in Figure

10.19c. Cut the rectangular pieces with prescribed colors as in Figure 10.33. Now rearrange the rectangular pieces to have the same order of colors. By overlapping the two congruent rectangles as before we obtain a common dissection as desired.

An important special case occurs when the parts themselves are the edges of two isoparametric n-gons, which may or may not be congruent. Then the foregoing dissection provides edge-to-edge conversion. In particular, if the n-gons are mirror images of one another, this gives another proof of Theorem 10.6.

We can also adapt designated dissections to polygons of equal area with unequal perimeters. Figure 10.34b shows an example, a square and equilateral triangle of equal areas with the shorter boundary of the square converted onto part of the longer boundary of the triangle, as indicated by the heavy lines. We can also divide the shorter boundary into pieces and convert them into designated pieces on the longer boundary, as was done in Figure 10.33. Moreover, we can select part of one boundary so dissection converts it into the interior of the other polygon. To do so, simply rotate by 180° (in the unfolded frame in Figure 10.33) the corresponding rectangle having that part as its base.

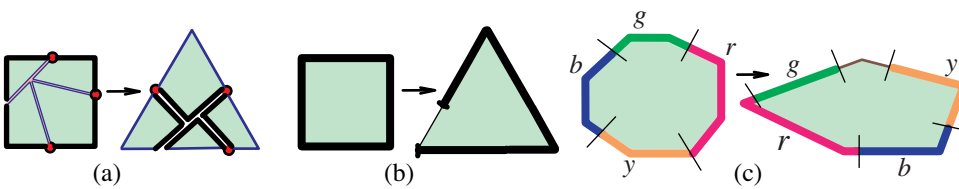


Figure 10.34: Polygonal regions with equal areas and unequal perimeters. In (a), the boundary of the square is converted to lie inside the triangle. In (b) and (c), parts of the shorter boundary are converted onto designated parts of the longer boundary.

10.16 CONCLUDING REMARKS

In view of Max Dehn's counterexample that settled Hilbert's third problem, extensions of our results on complete dissections in Part 2 to solids in 3-space are not always possible. Nevertheless, circumsolid shells, introduced in Chapter 4, have properties analogous to polygonal frames. For example, all circumsolid shells have constant thickness (see Theorem 4.19a). The analog of Lemma 10.1 is given by Theorem 4.11. If two polyhedral solids of equal volume and equal surface area can be dissected onto one another, it may be of interest to investigate whether a complete dissection exists that converts the boundaries onto one another.

Frederickson's book [41] gives an admirable introduction to the field of geometric dissections, and contains a valuable bibliography of results. Although there is a vast literature on standard polygonal dissection, we were not able to find any references relating to complete dissections of the type discussed in this chapter.

It is surprising that complete dissections have not been previously discussed in connection with cake slicing. Here's a natural question: Can we cut a cake with white icing on top and chocolate icing on its outer edges and rearrange the pieces

to form a cake of another shape (as in Figure 10.23) so that the white icing stays on top and the chocolate icing stays on the outer edges of the rearranged cake? Theorem 10.6 shows that this can always be done for polygonal cakes.

NOTES ON CHAPTER 10

The results in Part 1 are based on the authors' paper [13], which received a Lester Ford Award in 2005. Most of those in Part 2 appear in the authors' paper [28].

The ideas introduced in Part 1 suggest a host of complex and interesting problems for further study. First, we can find isoparametric rings such that the hole in each ring has a shape different from that of the outer contour. Although the results of this chapter deal with holes similar to the outer contour, they can also be extended to treat isoparametric holes that are dissimilar. Second, most of the results can be extended to higher-dimensional space, which allows many ways to extend isoparametric problems. For example, in 3-space we can compare volumes and surface areas, or we can compare surface areas and linear sizes such as edge-lengths and perimeters, or volumes and linear sizes. In higher-dimensional space we can compare n-dimensional and m-dimensional volumes and sizes. Third, instead of requiring that the perimeters and areas of two plane regions be equal, we could ask that they have prescribed ratios. This more general situation can be treated using the theorems of this chapter. There are similar extensions in higher-dimensional space.

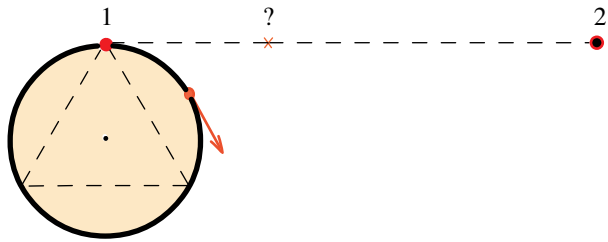
Chapter 11

ARCLENGTH AND TANVOLUTES

This problem can be easily solved by the methods developed in this chapter. The reader may wish to try solving it before reading the chapter.

The figure illustrates a special type of pursuit problem involving two missiles fired simultaneously. The problem is challenging because at first sight it appears that not enough information is given to solve it.

Missile 1 is initially at the point marked 1, and it moves clockwise along the circle at a constant speed v until it leaves the circle at some unknown point and continues at the same speed along the tangent line at that point.



Missile 2, initially at point marked 2, moves at constant speed $2v$, and is required to follow a path that intercepts missile 1, subject to the following constraints. Except for the initial moment when both missiles are fired, missile 2 cannot see missile 1, so this is a blind pursuit. The initial distance from point 1 to point 2 is equal to the perimeter of an equilateral triangle inscribed in the circle.

To allow for the possibility that missile 1 moves directly toward point 2, missile 2 travels along the straight line toward point 1 until it reaches the point marked ? where collision would occur. If collision occurs at ?, the pursuit is over. If not, this means that missile 1 detached from the circle at some other point (not known).

Find the path missile 2 *should follow to overtake missile* 1.

CONTENTS

PART 1: ARCLENGTH

PART 2: TANVOLUTES

11

As in Chapter 1, this chapter involves a plane base curve, called the tangency curve, along which a tangent vector moves continuously, its free end tracing another curve called the free-end curve. Earlier chapters dealt with the area of the region between the tangency curve and the free-end curve. Part 1 of this chapter relates the arclength functions of the two curves, and Part 2 introduces tanvolutes, a generalization of classical involutes.

As in Chapter 1, each curve is initially described by its geometric properties rather than by equations. The use of sweeping tangents determines the arclength as a function of the angle of turn of the tangent line. Examples in Part 1 include the arclength of a circle and tractrix, for which the tangent vector has constant length, and the arclength of exponential, parabolic, and cycloidal curves, for which the tangent vector has variable length. A knowledge of the arclength function gives a description of each curve by an intrinsic equation. Part 1 also treats the classical involute, a curve that intersects every tangent line to the base curve at a right angle. One example is the tractrix, an involute of a catenary.

Part 2 introduces the tanvolute, which intersects every tangent line to the base curve at any given fixed angle. This minor change in the definition of a classical concept leads to a wealth of new examples and phenomena. Our treatment is based on two differential equations relating four functions of the angle of turn: the arclength functions for the base curve and for its tanvolute obtained in Part 1, the length of the tangent from the base to the tanvolute, and the fixed angle. The parameters in the differential equations contribute many essential features to the solution curves. Even when the base curve is relatively simple, the variety in the shapes of the tanvolutes is remarkably rich. To illustrate, as a circle shrinks to a single point, its tanvolute becomes a logarithmic spiral!

An application is given to a generalized pursuit problem in which a missile is fired at constant speed in an unknown tangent direction from an unknown point on a base curve. Surprisingly, it can always be intercepted by a faster constant-speed missile that follows a specific tanvolute of a given base curve.

PART 1: ARCLENGTH

11.1 INTRODUCTION

In Chapter 1 we used sweeping tangents to calculate area. Now we use them to find arclength. Figure 11.1 shows a close-up view of the tangent sweep introduced in Figure 1.16. Each tangent segment is attached to a curve we call the *tangency curve* τ. The point of tangency moves along τ in a given direction, called the positive direction, as indicated by the arrowhead in Figure 11.1. At each point of τ the tangent line defines two rays, one in the positive direction of motion, the other in the opposite direction. It may be helpful to imagine an automobile driving along τ with its headlight beam indicating the direction of a tangent ray. If the automobile moves forward in the positive direction, its headlight beam indicates the direction of motion. If it drives backward, the headlight beam points in the direction opposite to the motion.

Assume a tangent vector moves continuously, always pointing in the positive direction during the motion, or else always pointing in the opposite direction. The moving tangent vector sweeps out a region we have called the *tangent sweep*. The free end of the tangent vector traces a curve σ called the *free-end curve*. There are two possible free-end curves, σ_+ generated when the tangent vectors point in the positive direction, and σ_- generated when the tangent vectors point in the backward direction.

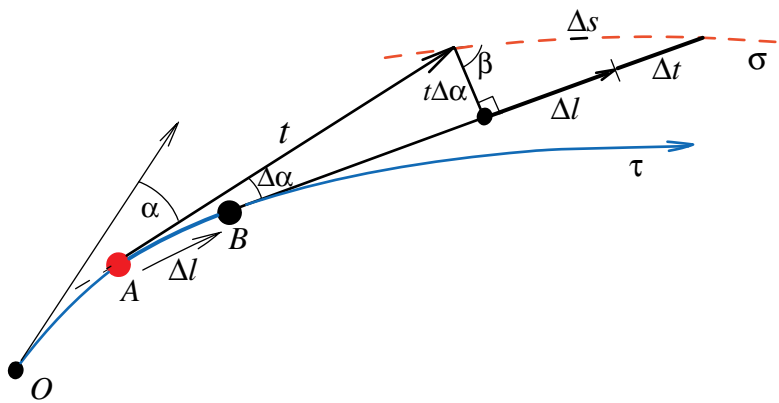


Figure 11.1: Arclength relations determined by two nearby tangents.

We denote by l the arclength function of τ, which tells us how far a particle moves along τ from an initial position to an arbitrary point on the curve. Similarly, s denotes the corresponding arclength function of σ. The arclength functions provide intrinsic descriptions of the curves. We denote by $t = t(\alpha)$ the function whose absolute value is the length of the tangent vector, where α is the angle between the moving tangent and an initial direction, for example, the tangent direction at a conveniently chosen point denoted by O in Figure 11.1. We know that $|t(\alpha)|$

provides a polar description of the tangent cluster, whose area is equal to that of the tangent sweep. Earlier chapters used knowledge of $t(\alpha)$ to determine areas of many classical regions. This chapter does not treat areas, but shows that a knowledge of $t(\alpha)$ gives the arclength functions l and s in terms of α.

11.2 ARCLENGTH OF TANGENCY CURVE: TANGENT IN THE FORWARD DIRECTION

Figure 11.1 shows the tangent segment of length $t = t(\alpha)$ at position A and at a nearby position B as α changes by $\Delta\alpha$. The point of tangency slides from A to B along τ through distance Δl, and the free end of the tangent vector traces a small arc Δs on σ. We treat arcs Δs and Δl as linear approximations to the curves, as depicted in Figure 11.1 for small $\Delta\alpha$. Thus, arc Δs is the hypotenuse of a right triangle, one leg of which is $t(\alpha)\Delta\alpha$, due to rotation of the tangent vector through angle $\Delta\alpha$ without changing its length. The other leg is made up of two parts, $\Delta l + \Delta t$, where Δl is caused by sliding the rotated tangent along τ without changing its length, and Δt is caused by the variability of its length. In what follows, we assume that arclength l increases as the point of tangency moves along τ, so Δl is positive. However t may increase or decrease during the motion, so Δt can be positive or negative. In Figure 11.1 the tangent vector points in the positive direction, and Δt is shown as positive.

For a given α, let β denote the complement of the angle between the tangents to τ and σ. Thus β is a function of α. The triangle ratio for $\tan\beta$ in Figure 11.1 yields the approximate relation

$$\Delta l + \Delta t \approx t(\alpha)\Delta\alpha \tan\beta.$$

Divide by $\Delta\alpha$ and let $\Delta\alpha \to 0$ to obtain a differential equation relating the derivatives of l and t:

$$\frac{dl}{d\alpha} + \frac{dt}{d\alpha} = t(\alpha)\tan\beta. \tag{11.1}$$

Integrating (11.1) over the interval $[0, \alpha]$ we find a natural equation for l in terms of t:

$$l(\alpha) - l(0) + t(\alpha) - t(0) = \int_0^\alpha t(\theta)\tan\beta\, d\theta. \tag{11.2}$$

In most applications we choose O so that $l(0) = 0$.

11.3 ARCLENGTH OF TANGENCY CURVE: TANGENT IN THE BACKWARD DIRECTION

In some applications it is convenient to reverse the direction of the tangent vector so it points opposite to the direction of increasing l, as shown in Figure 11.2b. Then the triangle ratio for $\tan\beta$ becomes $|\Delta l - \Delta t| \approx t(\alpha)\Delta\alpha \tan\beta$, where the absolute

value allows for the possibilities $\Delta l > \Delta t$ (Figure 11.2a), and $\Delta t > \Delta l$ (Figure 11.2b). This leads to a corresponding change in (11.1):

$$|\frac{dl}{d\alpha} - \frac{dt}{d\alpha}| = t(\alpha)\tan\beta, \tag{11.3}$$

and, instead of (11.2), we now have the backward integrated version

$$|l(\alpha) - l(0) - t(\alpha) + t(0)| = \int_0^\alpha t(\theta)\tan\beta\, d\theta. \tag{11.4}$$

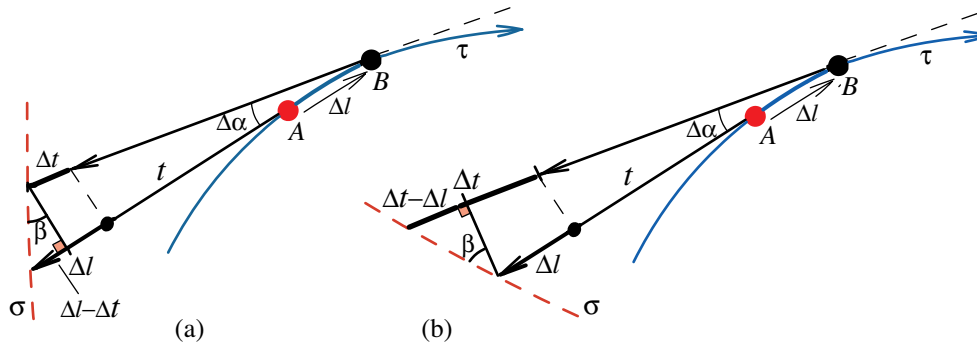


Figure 11.2: Arclength relations when tangent vector has backward direction.

Guided by the automobile analogy, we call (11.2) the *forward* relation and (11.4) the *backward* relation. They are identical when $t(\alpha)$ is constant.

Arclength of free-end curve.

Let Δs denote the change in arclength s of free-end curve σ (the hypotenuse of the small triangle in Figure 11.1), measured so that $s = 0$ when $\alpha = 0$. Then we have an approximate relation, which also holds for Figure 11.2:

$$\Delta s \approx \frac{t(\alpha)\Delta\alpha}{\cos\beta},$$

and which gives us a differential equation for s,

$$\frac{ds}{d\alpha} = \frac{t(\alpha)}{\cos\beta}. \tag{11.5}$$

Integrating (11.5), we obtain an equation for s in terms of t:

$$s(\alpha) = \int_0^\alpha \frac{t(\theta)}{\cos\beta}\, d\theta. \tag{11.6}$$

Relating arclengths of free-end curve and tangency curve.

A direct connection between the arclengths l of τ and s of σ can be found by eliminating β in the basic derivative relations (11.1) and (11.5). Square them and use the identity $\tan^2\beta = \sec^2\beta - 1$ to obtain the relation

$$(\frac{dl}{d\alpha} + \frac{dt}{d\alpha})^2 = (\frac{ds}{d\alpha})^2 - t^2(\alpha), \tag{11.7}$$

which involves arclength functions of τ and σ. It can also be derived directly by applying the Pythagorean Theorem to the triangle with hypotenuse Δs in Figure 11.1. Relation (11.7), in turn, gives an explicit formula expressing s in terms of l:

$$s(\alpha) = \int_0^\alpha \sqrt{t^2(\theta) + (\frac{dl}{d\theta} + \frac{dt}{d\theta})^2}\, d\theta. \tag{11.8}$$

It is easy to show that the classical arclength integral in polar coordinates is a limiting case of (11.8). Take τ to be a small circular arc that shrinks to a point. Then $t(\alpha)$ becomes the radial distance r from the point to σ, $dl/d\alpha \to 0$, and (11.8) becomes the classical arclength integral:

$$s(\alpha) = \int_0^\alpha \sqrt{r^2 + (\frac{dr}{d\theta})^2}\, d\theta.$$

Also, (11.8) can be transformed to resemble the classical integral in rectangular coordinates,

$$\int_a^b \sqrt{1 + (f'(x))^2}\, dx,$$

for the arclength of a curve $y = f(x)$ between two points $(a, f(a))$ and $(b, f(b))$. This follows from (11.8) by the substitution $dx = t(\theta)d\theta$ and $dy = dl + dt$. The orthogonality of dx and dy is revealed by the small right triangle in Figure 11.1.

One can also use (11.7) to formulate a counterpart to (11.8) expressing l in terms of s. This is left as an exercise for the reader.

11.4 EXAMPLES: TANGENTS OF CONSTANT LENGTH

All such examples are special cases of the bicycle problem discussed in Chapter 1, with constant length $t(\alpha) = k$. For any choice of τ and σ, both arclength formulas (11.2) and (11.4) become

$$l(\alpha) - l(0) = k \int_0^\alpha \tan\beta\, d\theta, \tag{11.9}$$

and (11.6) takes the form

$$s(\alpha) = k \int_0^\alpha \frac{1}{\cos\beta}\, d\theta. \tag{11.10}$$

Now we consider several examples by specializing τ and σ.

Circular arcs.

For arclengths of concentric circles of radii r and R, we use (11.9) and (11.10). From Figure 11.3a we find $\tan\beta = r/k$ and $\cos\beta = k/R$. Taking $l(0) = 0$, we obtain the familiar formulas for the length of a circular arc subtended by an angle α:

$$l(\alpha) = r\alpha, \quad s(\alpha) = R\alpha.$$

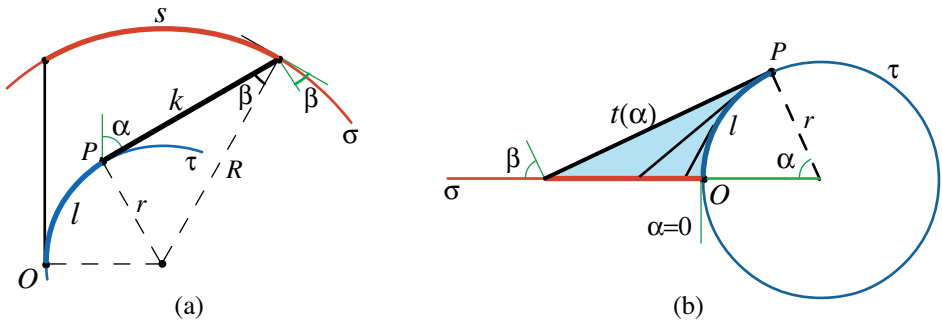


Figure 11.3: (a) Circular arclength. (b) Alternative choice of σ as a line through the center.

To demonstrate the flexibility in the choice of free-end curve we now choose σ to be a line through the center of the circle τ of radius r as in Figure 11.3b. In this example, $t(\alpha)$ is not constant, but $t(\alpha) = r\tan\alpha$ and $l(\alpha) = r\alpha$, so

$$t(\alpha) - l(\alpha) = r(\tan\alpha - \alpha).$$

On the other hand, from the backward relation (11.4) we also find (because now $\beta = \alpha$)

$$t(\alpha) - l(\alpha) = r\int_0^\alpha \tan^2\theta\, d\theta.$$

Therefore, as a fringe benefit we obtain the known integration formula

$$\int_0^\alpha \tan^2\theta\, d\theta = \tan\alpha - \alpha. \tag{11.11}$$

We can also derive (11.11) using the area of the tangent sweep in Figure 11.3b, which is that of a right triangle of edges r and $t(\alpha) = r\tan\alpha$, minus the area of the circular sector subtending angle α. But this area is also that of the tangent cluster which equals $\frac{1}{2}r^2\int_0^\alpha \tan^2\theta\, d\theta$, and again we obtain (11.11).

Tractrix.

Now take the tangency curve τ to be a *tractrix*, the trajectory of a toy on a taut string being pulled by a child walking along a linear path, which we take as σ, as

shown in Figure 1.18. We have shown in Chapter 1 that the area of the entire region between the tractrix and the horizontal line is that of a quarter of a circular disk.

Next we calculate the arclength. In Figure 11.4 the tangency curve τ is a tractrix, σ is the positive x axis, the tangent segment has constant length k, $\beta = \alpha$, and $l(0) = 0$, hence (11.9) becomes

$$l(\alpha) = k \int_0^\alpha \tan\theta \, d\theta = -k \log(\cos\alpha). \tag{11.12}$$

This is the natural equation for the tractrix, expressing its arclength in terms of α.

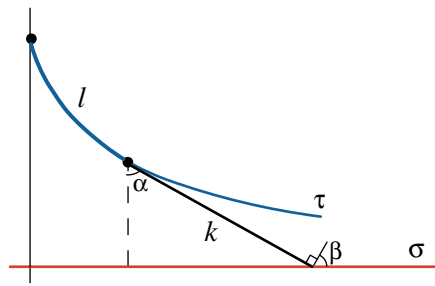


Figure 11.4: Arclength of tractrix.

If the point of tangency in Figure 11.4 has cartesian coordinates (x, y), then we have $\cos\alpha = y/k$, and (11.12) gives the arclength $L = L(y)$ of the tractrix from (0,k) to (x, y) as

$$L = -k \log \frac{y}{k}.$$

From this we find $y = ke^{-L/k}$, which means the ordinate y decreases exponentially with L. In particular, if an automobile drives along the tractrix with constant speed, its distance from the x axis decreases exponentially with time.

The arclength formula (11.10) for the free-end curve yields an unexpected fringe benefit for the tractrix. Using $\beta = \alpha$ in (11.10), we find

$$s(\alpha) = k \int_0^\alpha \frac{1}{\cos\theta} \, d\theta = k \log \frac{1 + \sin\alpha}{\cos\alpha}.$$

Let (x, y) be the point of tangency on the tractrix in Figure 11.4. Because σ is the x axis, we find

$$x = s(\alpha) - k\sin\alpha, \quad y = k\cos\alpha.$$

Using $\cos\alpha = y/k$, we find the classical cartesian representation of the tractrix as a direct consequence of natural equation (11.12):

$$x = k \log \frac{k + \sqrt{k^2 - y^2}}{y} - \sqrt{k^2 - y^2}.$$

In more general examples, where $t(\alpha)$ is not constant, whenever σ is a straight line we can take it to be the x axis and measure α from a vertical line. From a right triangle like that in Figure 11.4, the natural equation for s as a function of α yields

$$x = s(\alpha) - t(\alpha)\sin\alpha, \quad y = t(\alpha)\cos\alpha,$$

which are parametric equations of curve τ in rectangular coordinates.

11.5 TANGENT SEGMENTS OF VARIABLE LENGTH

Now we consider examples with tangent segments of variable length, and discuss arclengths of classical curves described by geometric properties rather than by equations. Our results for arclength provide natural or intrinsic equations for the curves.

Exponential.

In Figure 11.5 the tangency curve τ is the graph of an exponential $y = e^{x/b}$, where b is a positive constant, and the free-end curve σ is the x axis. It is known (see Chapter 1) that exponential curves are the only curves with constant subtangents. In fact, the exponential curve in Figure 11.5 has constant subtangents of length b, and we have used this geometric property to find the area of the region under an exponential curve without integral calculus.

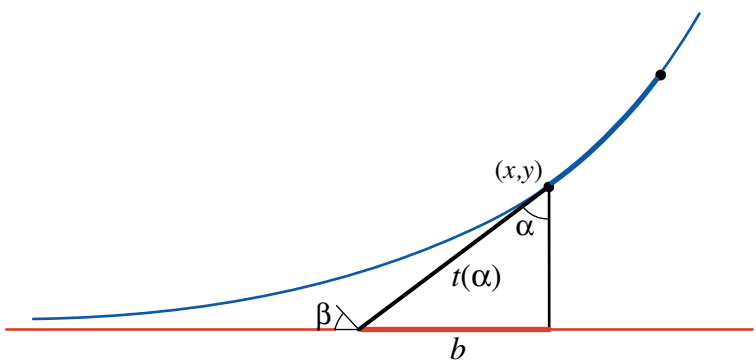


Figure 11.5: Arclength of exponential.

For arclength of the exponential, Figure 11.5 shows that $t(\alpha) = b/\sin\alpha$. Again we have $\beta = \alpha$, hence $t(\alpha)\tan\beta = b/\cos\alpha$. Integrate (11.4) over the interval $[\alpha_0, \alpha]$ to find, with $l(\alpha_0) = 0$,

$$l(\alpha) = b\left(\frac{1}{\sin\alpha_0} - \frac{1}{\sin\alpha} + \log\frac{1+\sin\alpha}{\cos\alpha} - \log\frac{1+\sin\alpha_0}{\cos\alpha_0}\right). \tag{11.13}$$

This natural equation for the exponential is valid for $0 < \alpha_0 < \alpha < \pi/2$.

At a general point of tangency (x, y) with angle α we have

$$t(\alpha) = \sqrt{b^2 + y^2}, \quad \sin\alpha = \frac{b}{\sqrt{b^2 + y^2}}, \quad \cos\alpha = \frac{y}{\sqrt{b^2 + y^2}},$$

and the natural equation (11.13) gives the classical formula for the arclength L of the exponential $y = e^{x/b}$ between (x_2, y_2) and (x_1, y_1), where $y_1 > y_2$:

$$L = \sqrt{b^2 + y_1^2} - \sqrt{b^2 + y_2^2} + b \log \frac{\sqrt{b^2 + y_2^2} + b}{y_2(\sqrt{b^2 + y_1^2} + b)}.$$

Parabola.

Figure 11.6a shows a portion of a parabola $y = x^2$ above the interval $[0, x]$. We take the parabola as tangency curve τ and the x axis as free-end curve σ. The tangent segment from τ at (x, x^2) to σ has subtangent $x/2$. From this property, we have already shown in two different ways (using Figures 11.6a and 11.6b) that the area of the parabolic segment between τ and σ is $x^3/3$. (See Chapter 1, where the general power function $y = x^n$ is also treated without integration.)

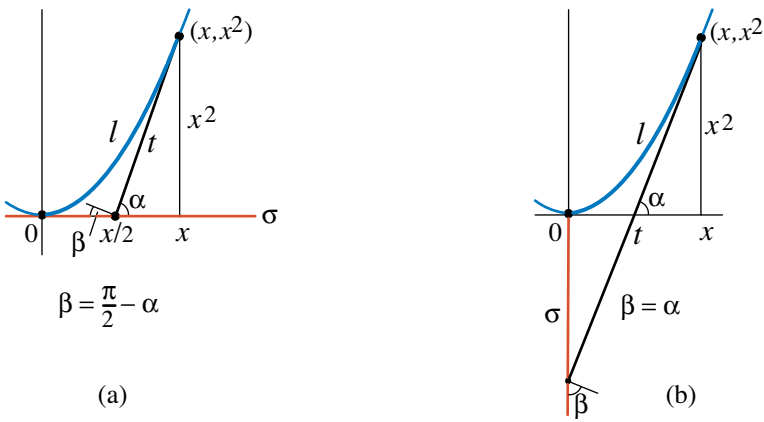


Figure 11.6: Two methods for calculating the arclength of a parabola.

Now we calculate the arclength traced by a point moving along the parabola from the origin to (x, x^2). To demonstrate the flexibility in the choice of free-end curve, Figures 11.6a and b illustrate two different ways for calculating the arclength. In both cases, α denotes the angle between the tangent at (x, x^2) and the x axis, but in Figure 11.6b the free-end curve σ is chosen to be the negative y axis.

In the backward formula (11.4) for arclength, $t(\alpha)\tan\beta$ appears in the integrand. To express $t(\alpha)\tan\beta$ in terms of α, note that in Figure 11.6a we have $\beta = \pi/2 - \alpha$, and hence $\tan\beta = \cot\alpha$. Also, $\tan\alpha = x^2/(x/2) = 4(x/2) = 4t(\alpha)\cos\alpha$, so that

$$t(\alpha) = \frac{\tan\alpha}{4\cos\alpha}, \quad \text{and} \quad t(\alpha)\tan\beta = \frac{1}{4\cos\alpha}.$$

Now we use (11.4) with $l(0) = t(0) = 0$ and we find that the arclength of the parabola is given by

$$l(\alpha) = t(\alpha) + \frac{1}{4}\int_0^\alpha \frac{d\theta}{\cos\theta} = \frac{1}{4}\left(\frac{\tan\alpha}{\cos\alpha} + \log\frac{1+\sin\alpha}{\cos\alpha}\right). \tag{11.14}$$

Figure 11.6b leads to the same formula for arclength. Now $\beta = \alpha$ and $t(\alpha)$ has twice the value in the foregoing calculation. In this case $t(\alpha) > l(\alpha)$ and a sign change is required in (11.14). We omit the details.

The arclength of a parabola was first found by Isaac Barrow, who used a different method and expressed the result in an equivalent form involving $\alpha/2$. Again, (11.14) represents the natural equation for the parabola.

In terms of x, the length $L(x)$ of the parabolic arc from the origin to (x, x^2) is

$$L(x) = |\frac{1}{2}x\sqrt{4x^2+1} + \frac{1}{4}\log(2x + \sqrt{4x^2+1})|.$$

Cycloid.

For the arclength of a cycloid, refer to Figure 11.7b. (Figure 11.7a was used in Chapter 2 to find the swept area.) The tangency curve τ is a cycloid, and the

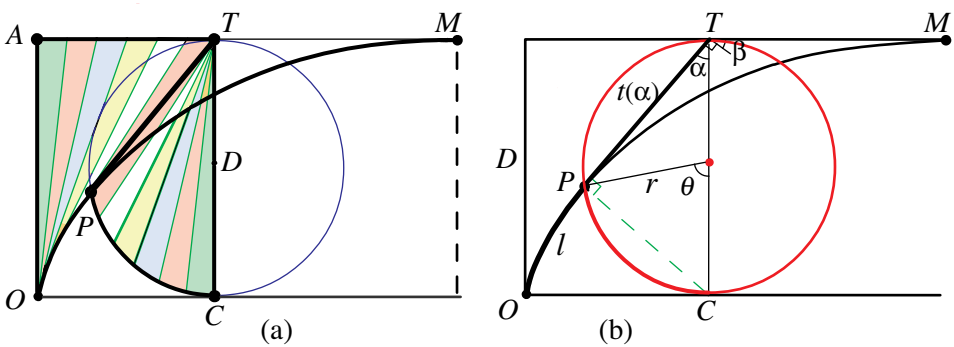


Figure 11.7: Length of cycloidal arc PM is twice that of tangent segment PT.

free-end curve is chosen as a horizontal line tangent to the highest point M of the cycloid, at distance D above the base. The tangent length $t(\alpha)$ is one leg of a right triangle inscribed in a semicircle of diameter D, so $t(\alpha) = D\cos\alpha$. Now $\beta = \alpha$ and $t(\alpha)\tan\beta = D\sin\alpha$, and (11.2) becomes

$$l(\alpha) = D - D\cos\alpha + D\int_0^\alpha \sin\theta\, d\theta,$$

which leads to the following simple formula for the arclength OP:

$$l(\alpha) = 2D(1-\cos\alpha) = 2D - 2t(\alpha). \tag{11.15}$$

In particular, when $\alpha = \pi/2$, arclength OM is $2D$, a result discovered by Christopher Wren. The general formula for $l(\alpha)$ implies that arclength PM, which is $2D - l(\alpha)$, is twice the length $t(\alpha)$ of tangent segment PT.

In Figure 11.7b, $D = 2r$, where r is the radius of the rolling disk that traces the cycloid. Denote by $L(\theta)$ the length of cycloidal arc OP in terms of the angle of turn θ of the rolling disk. Then the formula for arclength OP becomes

$$L(\theta) = 4r(1 - \cos\frac{\theta}{2}). \tag{11.16}$$

Epicycloid and hypocycloid.

For the arclength of an epicycloid, refer to Figure 2.10 in Chapter 2. In this case, $\alpha = \kappa\beta$, where $\kappa = 1 + 2r/R$. In Figure 2.10 the tangency curve τ is the epicycloid and the free-end curve σ is the circular arc of radius $R + 2r$. We see that $PT = t(\alpha) = 2r\cos\beta$ and the integrand in (11.2), when expressed in terms of β, becomes $2r\kappa\sin\beta\, d\beta$. Now $\beta = \theta/2$, where θ is the angle of turn of the rolling disk of radius r. In terms of θ we find that the length $L(\theta)$ of the epicycloidal arc OP in Figure 2.10 is

$$L(\theta) = 4r(1 + \frac{r}{R})(1 - \cos\frac{\theta}{2}). \tag{11.17}$$

For a *hypocycloid*, in which the circle of radius r rolls inside the circumference of the fixed circle of radius R as in Figure 2.11, the formula for arclength is similar to (11.17), with the factor $(1+r/R)$ replaced by $(1-r/R)$. When $R = \infty$ both reduce to (11.16) for a cycloid.

It is of interest to note that the sum of the arclength of an epicycloid and of its complementary hypocycloid is twice that of the cycloidal arc in (11.16). This can be regarded as a limiting case of the corresponding property of complementary trochogonal curves obtained in Section 3.14 by an elementary method.

11.6 CLASSICAL INVOLUTE AND EVOLUTE

Consider a family of normals to a given curve σ, as in Figure 11.8. Their envelope τ is called the *evolute* of σ, and σ, in turn, is called the *involute* of τ. We choose τ as tangency curve and σ as free-end curve. Because $\beta = 0$, the integral in (11.4) vanishes and we find

$$l(\alpha) - l(0) = t(\alpha) - t(0). \tag{11.18}$$

If we take $l(0) = t(0) = 0$, (11.18) becomes

$$l(\alpha) = t(\alpha). \tag{11.19}$$

This verifies the intuitively apparent fact that if a taut inelastic string is unwrapped from a point O on τ, its free end will trace the involute σ.

Because $t(\alpha) = l(\alpha)$, the area $A(\alpha)$ of the region between the evolute and involute swept by the tangent as it moves from O to P is

$$A(\alpha) = \frac{1}{2}\int_0^\alpha l^2(\theta)d\theta. \tag{11.20}$$

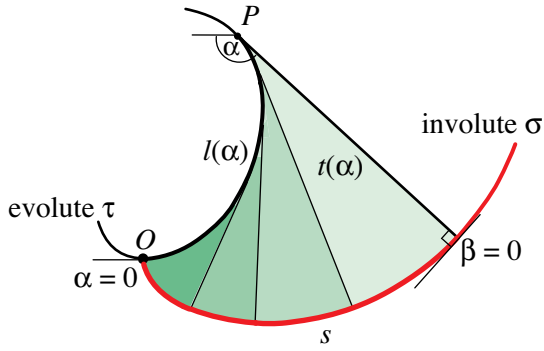


Figure 11.8: Classical involute-evolute relations.

Formula (11.6) for the free-end curve gives the arclength of the involute:

$$s(\alpha) = \int_0^\alpha l(\theta)d\theta. \tag{11.21}$$

From (11.21) we can find the arclength l of the evolute τ in terms of s by differentiation with respect to α:

$$l(\alpha) = s'(\alpha). \tag{11.22}$$

Thus, the traditional involute-evolute relations (11.18), (11.20), and (11.21) are merely special cases of our basic relations when $t(\alpha)$ is chosen to be $l(\alpha)$, the arclength of the tangency curve.

Involute of a circle.

Figure 11.9a shows the special case in which the tangency curve τ is a circle of radius r. Here $t(\alpha) = r\alpha$, where α is measured counterclockwise with $\alpha = 0$ at O. Now use (11.21) with $l(\alpha) = t(\alpha) = r\alpha$ to obtain $s(\alpha) = \frac{1}{2}r\alpha^2$. This intrinsic equation expresses the arclength of the involute of a circle in terms of the unwrapping angle. Figure 11.9b shows the tangent cluster of the tangent sweep in Figure 11.9a. Its boundary is an Archimedean spiral because the polar radius is proportional to angle α. Let $A(\alpha)$ denote the area of the region swept by the polar radius $r\theta$ of an Archimedean spiral as θ varies from 0 to α. Formula (11.20) for the area swept between the circle and its involute implies

$$A(\alpha) = \frac{1}{2}\int_0^\alpha (r\theta)^2\, d\theta = \frac{1}{6}r^2\alpha^3. \tag{11.23}$$

This can be written as $\frac{1}{3}(R^2\alpha/2)$, where $R = r\alpha$. This result, which was found by Archimedes, states that the area of the region swept by the polar radius of an Archimedean spiral is one-third the area of the circular sector whose radius is the final polar radius of the spiral.

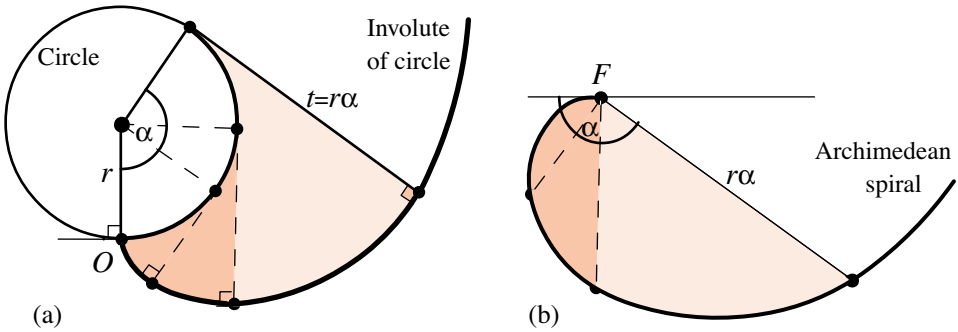


Figure 11.9: (a) Tangent sweep between a circle and its involute. (b) Its tangent cluster is bounded by an Archimedean spiral.

The formulas for the arclength and area of the involute of a circle were obtained differently in Chapter 3 in (3.46) and (3.47).

Evolute of a tractrix.

Figure 11.10a shows a tractrix and an arc OP of its evolute. If a string lying along the evolute is unwrapped from O, its free end traces a portion of a tractrix joining point O to point T. The tangent segment from the tractrix at T to the base line AX has constant length k, which is equal to the height of O above AX. We shall determine the length l of arc OP, and the area of the ordinate set $OAXP$ between the evolute and the base AX.

According to (11.22), $l(\alpha) = s'(\alpha)$, where s is the arclength of the involute, the tractrix in this case, which we have calculated in (11.12). Renaming l in (11.12) as s we find, by differentiation, $s'(\alpha) = k \tan \alpha$, so arclength $l = l(\alpha)$ of the evolute is

$$l(\alpha) = k \tan \alpha. \tag{11.24}$$

Next we show that the area of ordinate set $OAXP$ is equal to that of rectangle $TPT'X$ in Figure 11.10b, with vertical diagonal XP and edges of lengths k and l. Region $OAXP$ consists of three parts: OAX swept by tangent segments of length k to the tractrix, OTP swept by tangent segments of variable length $l(\alpha)$ to the evolute OP, and triangle PTX.

Divide triangle $PT'X$ into two regions, a circular sector $PA'T'$, and region $A'XT'$. Sector $PA'T'$ is the tangent cluster of OAX, the tangent sweep of the tractrix, so they have equal areas. Region $A'XT'$ is swept by tangent segments to the circular arc $A'T'$, which are obtained from tangent sweep OTP by parallel translation of each tangent segment from the evolute. Therefore both swept regions $A'T'X$ and OTP have the same tangent cluster, so they have equal areas.

Thus, the sum of the areas of the two shaded regions on the left of the diagonal PX is equal to the area of the shaded triangle $PT'X$ on the other side of the

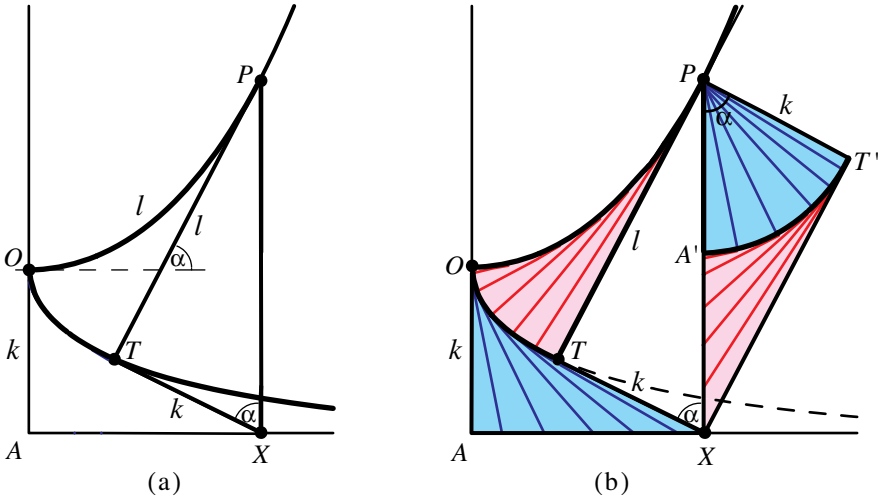


Figure 11.10: (a) The evolute of a tractrix. (b) Proof that area of the ordinate set $OAXP$ is that of rectangle $TPT'X$. This implies that the evolute of a tractrix is a catenary.

diagonal. By adding the area of the common unshaded triangle PTX we find

$$\text{area of ordinate set}\, OAXP = \text{area of rectangle}\, TPT'X = kl(\alpha), \tag{11.25}$$

where $l(\alpha)$ is given by (11.24).

The same formulas (11.24) and (11.25) were obtained differently in (3.51) and (3.50) in Chapter 3 as the arclength and area formulas for a parabolic catenary.

Catenary as evolute of a tractrix.

Now we use our results for arclength and area to deduce the known fact that the evolute of a tractrix is a catenary, the shape of a uniform flexible chain that hangs under its own weight. Arclength formula (11.24) and area formula (11.25) together show that

$$\int_0^x y(u)\,du = k^2 y',$$

which, when differentiated, gives $y = k^2 y''$. The unique solution to the differential equation with $y(0) = k$ and $y'(0) = 0$ is

$$y = k\cosh\frac{x}{k} = k\frac{e^{x/k} + e^{-x/k}}{2}. \tag{11.26}$$

This is the cartesian equation for the catenary, and (11.24) is its natural equation.

Formulas (11.24) and (11.25) for arclength and area can also be expressed in terms of hyperbolic functions. Let $L(x)$ denote the arclength of the catenary from

(0,k) to (x, y), and let $A(x)$ denote the area of the corresponding ordinate set. Then (11.24) and (11.25) give us the classical formulas

$$L(x) = \sinh \frac{x}{k}, \text{ and } A(x) = k^2 \sinh \frac{x}{k},$$

where $\sinh x = (e^x - e^{-x})/2$ is the derivative of $\cosh x$.

Also, (11.25) tells us that the area of the ordinate set between the catenary and the interval AX is equal to the arclength $l(\alpha)$ multiplied by the height k of the catenary's lowest point above AX, and, by (11.24), it is also equal to k^2 times the slope of the tangent line at P. Moreover, the ordinate XP of the catenary is equal to $k/\cos\alpha$.

The catenary is well known as the shape of a uniform flexible chain that hangs under its own weight. The standard proof of this makes use of a triangle of equilibrium of forces that is similar to triangle PTX in Figure 11.10a.

We can derive the arclength of the tractrix in (11.12) from the natural equation of the catenary in (11.24). Choose again the tractrix as the free-end curve with the catenary as tangency curve, and use (11.6), taking $\beta = 0$ and $t(\alpha) = l(\alpha) = k \tan\alpha$.

Generalized pursuit curve.

Figure 11.11a shows a tangency curve τ and a horizontal free-end curve σ. At a general point of τ a tangent segment of length $t(\alpha)$ cuts off a subtangent of length $b(\alpha)$. For a tractrix, $t(\alpha)$ is constant, and for an exponential, $b(\alpha)$ is constant. Now we consider the more general case, treated earlier in Section 1.17, in which a convex combination of $t(\alpha)$ and $b(\alpha)$ is constant, say

$$\mu t(\alpha) + \nu b(\alpha) = C, \tag{11.27}$$

for some choice of nonnegative μ and ν, with $\mu + \nu = 1$. If $\nu = 0$, $\tau(\alpha)$ is constant and the tangency curve τ is a tractrix, which can be regarded as a pursuit curve. If $\mu = \nu$ then $t(\alpha) + b(\alpha)$ is constant, and τ is another pursuit curve in which a fox running on σ is pursued by a dog on τ having the same speed as the fox. Because of these examples, we refer to any curve satisfying (11.27) as a *generalized pursuit curve*.

Now we will show once more that the tangent cluster of a generalized pursuit curve is bounded by a conic section with eccentricity ν/μ and a focus at the common point F of the translated segments. An example is shown in Figure 11.11b.

In Figure 11.11a we have $\beta = \alpha$ and $b(\alpha) = t(\alpha)\sin\alpha$. Let $D = t(0)$ and let $e = \nu/\mu$, where $\mu \neq 0$. Then $b(0) = 0$ and (11.27) implies

$$t(\alpha) = \frac{D}{1 + e\sin\alpha}. \tag{11.28}$$

This is the polar equation with radial distance $t(\alpha)$ of a conic with eccentricity e and focus at F.

Thus, the area of the tangent sweep in Figure 11.11a is equal to that of the corresponding tangent cluster in Figure 11.11b, a sector of a conic section swept by

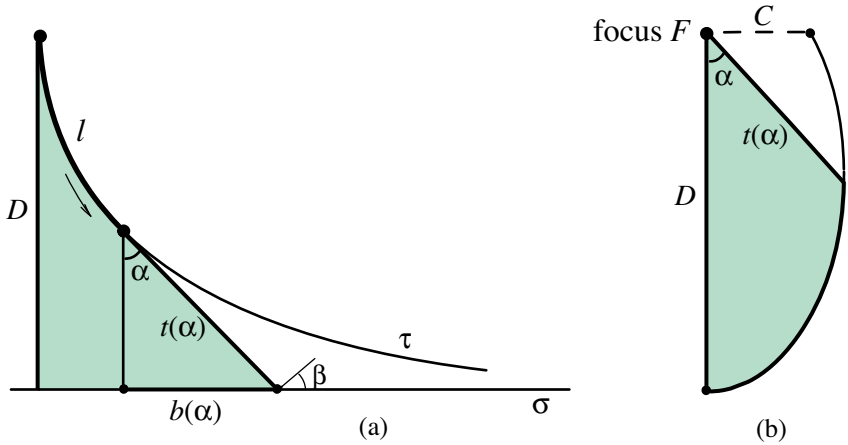


Figure 11.11: (a) Generalized pursuit curve: $\mu t(\alpha) + \nu b(\alpha) = C$. (b) Tangent cluster of tangent sweep in (a) is a focal sector of a conic section.

a focal radius. This is analogous to the Keplerian sector swept by the radius vector from the sun to an orbiting planet. We have found the area of the shaded region below the pursuit curve without knowing any equation that describes the curve. Its natural equation will be derived in (11.30) and (11.31).

As Figure 11.11a suggests, the distances $t(\alpha)$ and $b(\alpha)$ are asymptotically equal as $\alpha \to \pi/2$. From (11.27) and (11.28) we find that the asymptotic pursuit distance is $t(\pi/2) = C = D\mu$.

Now we determine the arclength l of τ from the forward formula (11.2). Using (11.28) and $t(0) = D$ in (11.2) we find

$$l(\alpha) = \frac{De\sin\alpha}{1 + e\sin\alpha} + D\int_0^\alpha \frac{\sin\theta}{(1 + e\sin\theta)(\cos\theta)}\, d\theta. \tag{11.29}$$

The substitution $u = \sin\theta$ converts the integral in (11.29) into

$$\int_0^{\sin\alpha} \frac{u}{(1 + eu)(1 - u^2)}\, du.$$

Partial fraction decomposition requires us to consider two cases, $e = 1$ and $e \neq 1$.

Case $e = 1$

In this case the integrand is given by

$$\frac{1}{4}\left(\frac{1}{1+u} - \frac{2}{(1+u)^2} + \frac{1}{1-u}\right),$$

and (11.29) yields the intrinsic equation for the arclength of a classical pursuit curve:

$$l(\alpha) = \frac{D\sin\alpha}{2(1 + \sin\alpha)} + \frac{D}{4}\log\frac{1 + \sin\alpha}{1 - \sin\alpha}. \tag{11.30}$$

In this case, using (11.6), we find that $s(\alpha)$ is equal to $l(\alpha)$ as given by (11.30). This is to be expected because the dog and fox have equal speeds.

Case $e \neq 1$

In this case (11.29) leads to an intrinsic equation for the arclength of a generalized pursuit curve:

$$l(\alpha) = \frac{De\sin\alpha}{1+e\sin\alpha} + \frac{D}{2}\left(\frac{1}{e-1}\log\frac{1+\sin\alpha}{1+e\sin\alpha} - \frac{1}{e+1}\log\frac{1-\sin\alpha}{1+e\sin\alpha}\right). \qquad (11.31)$$

It can be shown (as expected) that (11.30) is a limiting case of (11.31) as $e \to 1$.

The exponential curve corresponds to $\mu = 0$ in (11.27), or the limiting case $e \to \infty$. By considering $l(\alpha) - l(\alpha_0)$ in (11.31) and letting $e \to \infty$ in such a way that $D/e \to b$, (where b is the constant subtangent of the exponential) we are led once more to (11.13). For $e > 1$ the conic (11.28) is a hyperbola, which, when $e \to \infty$, degenerates into a line at distance b from the focus. This line appears as the dashed vertical line in the cluster triangle in Figure 1.23, where the common cluster point F is at distance b from the line and serves as the focus of the degenerate hyperbola.

In the generalized pursuit curve (11.27), σ is the x axis and $\beta = \alpha$. Hence, as mentioned earlier, we can use (11.6) to calculate $s(\alpha)$ and obtain parametric equations for the rectangular coordinates of any point on the pursuit curve τ.

PART 2: TANVOLUTES

11.7 TANVOLUTES

This chapter began with a tangency curve τ and a tangent vector of length $|t(\alpha)|$ whose endpoint traces a free-end curve σ as the vector moves along τ at tangency angle α. The arclength functions $l(\alpha)$ for τ and $s(\alpha)$ for σ are connected by basic equations (11.3) and (11.5), in which the parameter β (also a function of α) represents the complement of the angle between the tangents to τ and σ.

The classical involute-evolute relations in Section 11.6 came from treating the special case $\beta = 0$ and $t(\alpha) = l(\alpha)$. The involute σ is traced by the free end of a taut inelastic string unwrapped from tangency curve τ so that it travels along the tangent to σ. Figure 11.12a shows a classical involute-evolute pair (σ, τ).

In the rest of this chapter we unwrap an elastic string, as suggested in Figure 11.12b, by changing its length in such a way that its free end moves at a constant angle β with the normal to the string, which is tangent to the involute through that point. We refer to β as the *angle of attack*, terminology borrowed from aerodynamics. In general, β can vary from point to point, but when β is constant we call the resulting free-end curve, which depends on β, a *tanvolute*, or more specifically, a *β-tanvolute*, and we denote it by σ_β. The word "tanvolute" is a blending of tangent and volute.

It suffices to assume that $-\frac{\pi}{2} < \beta < \frac{\pi}{2}$. When $\beta = 0$, the tanvolute σ_0 is the classical involute of τ. As β increases from $-\pi/2$ to 0, the tanvolutes form a

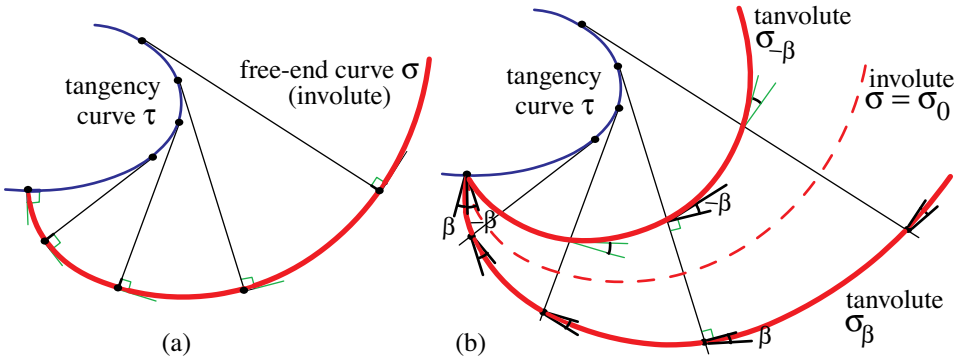


Figure 11.12: (a) Classical involute of τ. (b) Tanvolute of τ formed by unwrapping an elastic string from τ so that its free end moves at a constant angle of attack β.

one-parameter family intermediate to the tangency curve and its involute σ_0. As β continues to increase from 0 to $\pi/2$ the tanvolutes form another one-parameter family of curves expanding outward beyond the involute as indicated by the example in Figure 11.12b.

Figure 11.12a shows a classical involute of τ, and Figure 11.12b shows two β-tanvolutes of τ corresponding to two values of β of opposite sign.

Figure 11.13 shows examples of tanvolutes that occur in a well-known pursuit problem. Here n ants start at the vertices of a regular n-gon, each ant pursuing

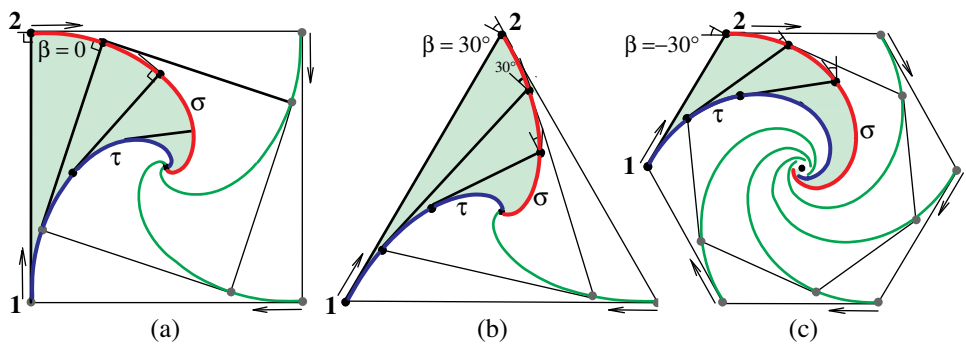


Figure 11.13: Pursuit problem with n ants at the vertices of a regular n-gon. While ant 1 pursues ant 2, ant 2 traces the β-tanvolute of the path of ant 1.

its nearest clockwise neighbor at the same speed. At any time, the ants are at the vertices of a smaller similar rotated n-gon, shown for $n = 4, 3$, and 6 in Figure 11.13. In each case, ant 1 traces a (logarithmic) spiral τ toward the center, while its neighbor, ant 2, traces the β-tanvolute of τ with constant β (by similarity), given by $\beta = \frac{2\pi}{n} - \frac{\pi}{2}$ radians. A new application to pursuit problems is in Section 11.13.

As we show later, a logarithmic spiral can be regarded as the tanvolute of a single point. Thus, tanvolutes also generalize the logarithmic spiral with its pole replaced by a curve.

11.8 BASIC FUNCTIONS AND THREE BASIC PROBLEMS

In the rest of this chapter, we consider a curve τ (the tangency curve), together with a moving tangent line. The moving tangent point on τ is described by a position vector $\mathbf{x}$, as indicated in Figure 11.14. Denote by $l = l(\alpha)$ the arclength function of τ, where α is the angle, measured counterclockwise, between the moving tangent and some initial direction, chosen so that $l(0) = 0$. As the point of tangency moves continuously along τ, we assume that the tangent vector angle α changes monotonically, increasing when the tangent turns counterclockwise. As α varies from $-\infty$ to $+\infty$, the arclength function can be positive or negative. The absolute value $|l(\alpha)|$ tells us how far the point of tangency moves along τ when the turning angle changes from 0 to α.

The derivative $d\mathbf{x}/dl = \mathbf{T}$, the unit tangent vector. The derivative of $\mathbf{T}$, in turn, is given by $d\mathbf{T}/dl = \kappa\mathbf{N}$, where $\kappa = d\alpha/dl$ is the curvature, and $\mathbf{N}$ is the principal unit normal, as in Figure 11.14.

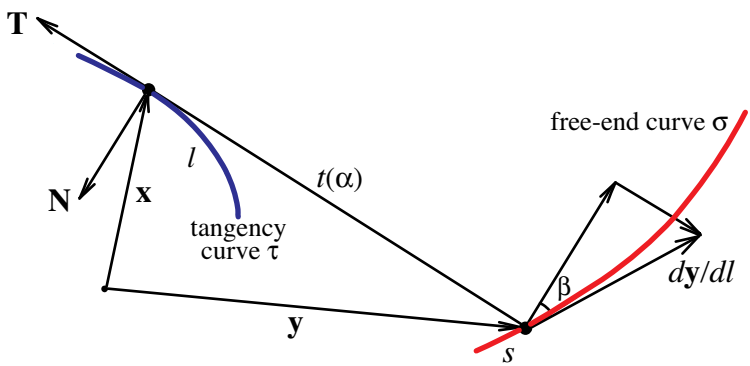


Figure 11.14: Tangency curve τ described by position vector $\mathbf{x}$, and free-end curve σ described by position vector $\mathbf{y} = \mathbf{x} - t\mathbf{T}$, where $\mathbf{T}$ is the unit tangent vector of τ.

Denote by $t = t(\alpha)$ the function whose absolute value is the length of the tangent vector from τ to the free-end curve σ. The vector $\mathbf{y} = \mathbf{x} - t\mathbf{T}$ is the position vector of the curve σ, as shown in Figure 11.14. Denote by $s = s(\alpha)$ the corresponding arclength function of σ with respect to angle α, measured so that $s(0) = 0$. Unlike $l(\alpha)$, which gives an intrinsic description of τ, the function $s(\alpha)$ does not always provide an intrinsic description of σ because α is not necessarily the angle swept by the tangent to σ. In this chapter we assume that the functions l, t, and s are continuously differentiable.

At each point where the tangent intersects σ, let β denote the angle of attack as described earlier, with $|\beta| < \pi/2$. For tanvolutes, β is constant. But in Sections

11.9 and 11.19, we allow β to vary from point to point and regard β as a function of α.

We relate the four functions β, t, l, and s through two differential equations in Theorem 11.1. When β is constant, we treat three problems in which we specify one of t, l, s and determine the other two.

11.9 BASIC DIFFERENTIAL EQUATIONS

Theorem 11.1. *The four functions l, s, t, and β are related by the differential equations*

$$\frac{dt}{d\alpha} - \frac{dl}{d\alpha} = t\tan\beta, \tag{11.32}$$

and

$$\frac{ds}{d\alpha} = \frac{t}{\cos\beta}. \tag{11.33}$$

This theorem was proved in Section 11.3 by a graphical method using linear approximations to the curves. Here we give a brief analytic argument.

Proof. Refer to Figure 11.14, where $\mathbf{x}$ is the position vector of the moving point of tangency on τ, and $\mathbf{y} = \mathbf{x} - t\mathbf{T}$ is the corresponding position vector of the free-end curve σ. Differentiation with respect to l gives

$$\frac{d\mathbf{y}}{dl} = (1 - \frac{dt}{dl})\mathbf{T} - t\kappa\mathbf{N} = \kappa[(\frac{dl}{d\alpha} - \frac{dt}{d\alpha})\mathbf{T} - t\mathbf{N}]. \tag{11.34}$$

The vector $d\mathbf{y}/dl$, shown tangent to σ in Figure 11.14, is the sum of two perpendicular vectors, one parallel to $\mathbf{T}$ and another parallel to $\mathbf{N}$ in the directions indicated. The three vectors form a right triangle, with angle β adjacent to the hypotenuse. From this triangle and (11.34) we deduce (11.32). Because the length of $d\mathbf{y}/dl$ is ds/dl, the same triangle shows that $(ds/dl)\cos\beta = \kappa t$. But $ds/dl = \kappa ds/d\alpha$, and we obtain (11.33).

11.10 CONSTANT ANGLE OF ATTACK: β-TANVOLUTES.

In general, β can vary with α, but for tanvolutes we keep β constant. Then as the tangent to τ sweeps through angle α, the corresponding tangent to the β-tanvolute sweeps through the same angle, as shown in Figure 11.15. This geometric property plays an important role in the analysis because, for constant β, the arclength s yields an intrinsic description of σ_β.

For constant β, we introduce constants c and γ, where

$$c = \tan\beta \text{ and } \gamma = \frac{1}{\cos\beta}.$$

Then the differential equations (11.32) and (11.33) become linear in t, l, and s:

$$\frac{dt}{d\alpha} - \frac{dl}{d\alpha} = ct, \text{ and } \frac{ds}{d\alpha} = \gamma t.$$

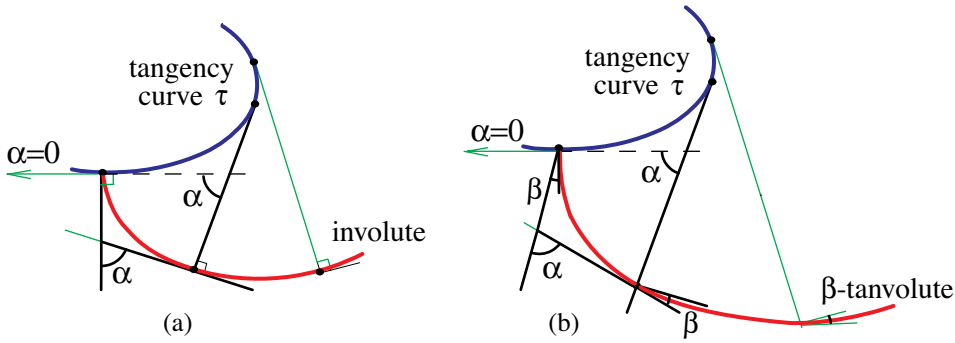


Figure 11.15: For constant β, when the tangent to τ sweeps through angle α, the corresponding tangent to the β-tanvolute sweeps through the same angle α.

As mentioned earlier, we treat three problems, in which we specify one of the functions and determine the other two. The first specifies l and determines t and s, which we denote as t_β and s_β. The other two are in Sections 11.14 and 11.15.

11.11 PROBLEM 1: FINDING β-TANVOLUTES OF A GIVEN CURVE

Problem 1. *For a constant β, a function l with $l(0) = 0$, and a real T, determine t_β and s_β satisfying* (11.32) *and* (11.33) *with initial conditions*

$$t_\beta(0) = T, \quad s_\beta(0) = 0.$$

Solution. The differential equation (11.32) is first-order linear for t_β with solution

$$t_\beta(\alpha) = l(\alpha) + Te^{c\alpha} + ce^{c\alpha}\int_0^\alpha l(\theta)e^{-c\theta}d\theta. \tag{11.35}$$

(We use the notation c and γ from the foregoing section.) Once t_β is known we can determine s_β by integrating γt_β, as suggested by (11.33). However, to avoid integrating an integral, we determine s_β differently. First use (11.33) to replace $t_\beta(\alpha)$ on the right of (11.32) and obtain the following version of (11.32):

$$\frac{dt_\beta}{d\alpha} = \frac{dl}{d\alpha} + \frac{c}{\gamma}\frac{ds_\beta}{d\alpha}.$$

Integrate this over the interval $[0, \alpha]$ and use the initial conditions to get

$$t_\beta(\alpha) - T = l(\alpha) + \frac{c}{\gamma}s_\beta(\alpha). \tag{11.36}$$

If $\beta = 0$, then $c = 0, \gamma = 1$, (11.36) becomes $t_0(\alpha) = T + l(\alpha)$, and $s_0(\alpha)$ is obtained by integration:

$$s_0(\alpha) = T\alpha + \int_0^\alpha l(\theta)d\theta. \tag{11.37}$$

If $\beta \neq 0$, then $c \neq 0$ and we can solve for $s_\beta(\alpha)$ in (11.36), and use (11.35) to obtain

$$s_\beta(\alpha) = \gamma T \frac{e^{c\alpha} - 1}{c} + \gamma e^{c\alpha} \int_0^\alpha l(\theta) e^{-c\theta} d\theta. \tag{11.38}$$

When $c \to 0$ in (11.38) we get (11.37) as a limiting case. Thus, (11.35) and (11.38) provide the solution to Problem 1.

The rightmost term in each of (11.38) and (11.35) contains the convolution integral

$$I(\alpha) = \int_0^\alpha l(\theta) e^{c(\alpha-\theta)} d\theta, \tag{11.39}$$

hence (11.38) and (11.35) can also be written, respectively, as

$$s_\beta(\alpha) = \gamma T \frac{e^{c\alpha} - 1}{c} + \gamma I(\alpha), \tag{11.40}$$

and

$$t_\beta(\alpha) = l(\alpha) + T e^{c\alpha} + c I(\alpha). \tag{11.41}$$

When $\alpha = 0$, these give us $s_\beta(0) = 0$ and $t_\beta(0) = T$, as expected. The free endpoint of the string starts at distance $|T|$ from the tangency curve. We always take $\alpha = 0$ in the direction of the initial tangent to the tanvolute.

Figure 11.16 shows the shape of β-tanvolutes for values of β near the extremities of the interval $[-\pi/2, \pi/2]$. In Figure 11.16a, β is negative and nearly $-\pi/2$, while

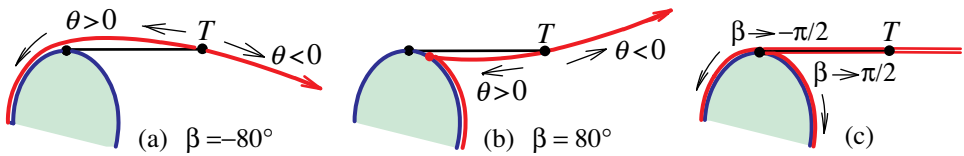


Figure 11.16: Behavior of β-tanvolute when β is near the extreme values $\pm\pi/2$.

in Figure 11.16b, β is near $\pi/2$. Figure 11.16c shows the limiting behavior as $\beta \to \pm\pi/2$; the tanvolute becomes a tangent ray to τ and then wraps around τ.

11.12 EXAMPLES ILLUSTRATING PROBLEM 1

Next we consider examples for which the convolution integral $I(\alpha)$ in (11.39) is easily evaluated. We begin with a circle, whose arclength function is linear in α.

Example 1 (Tanvolutes of a circle). For a circle of radius r the arclength is $l(\alpha) = r\alpha$, and (11.39) becomes

$$I(\alpha) = r e^{c\alpha} \int_0^\alpha \theta e^{-c\theta} d\theta = r \frac{e^{c\alpha} - 1}{c^2} - \frac{r\alpha}{c},$$

which, when used in (11.40) and (11.41), gives us the formulas

$$s_\beta(\alpha) = \gamma(T + \frac{r}{c}) \frac{e^{c\alpha} - 1}{c} - \frac{\gamma r \alpha}{c}, \tag{11.42}$$

$$t_\beta(\alpha) = Te^{c\alpha} + r\frac{e^{c\alpha}-1}{c}. \tag{11.43}$$

Classical case: Involute of a circle.

When $\beta = 0$, (11.36) and (11.37) yield

$$t_0(\alpha) = T + r\alpha, \text{ and } s_0(\alpha) = T\alpha + \frac{1}{2}r\alpha^2. \tag{11.44}$$

Figure 11.17 shows how the involute of a circle depends on T.

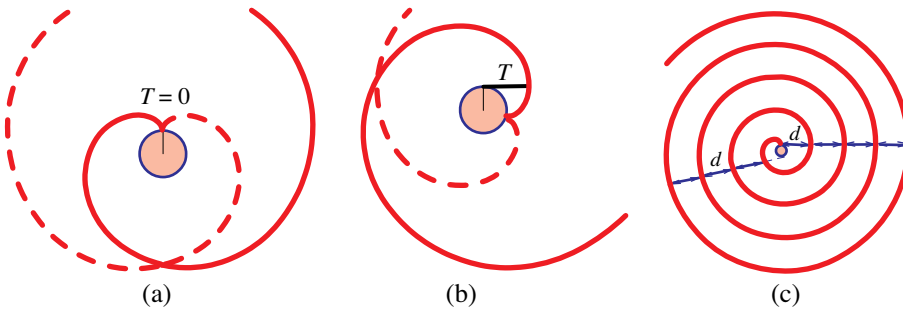


Figure 11.17: Involutes of a circle depending on T. (a) When $T = 0$ the involute has two branches spiraling outward symmetrically from the point of contact on the circle. (b) When $T > 0$, the involute is a rotated version of that in (a), spiraling outward from a different point on the circle. (c) Bird's eye view showing one of the spirals with constant spacing between its arms.

Special tanvolute: No exponential influence.

In (11.42) and (11.43), the exponential term $e^{c\alpha}$ disappears when $T = -r/c$. For this value of T, which we label T_*, we denote the arclength function in (11.42) by s^*_β and the corresponding tanvolute by σ^*_β, which we call a *special tanvolute*. Because $c/\gamma = \sin\beta$, from (11.42) we find $s^*_\beta(\alpha) = R_\beta\alpha$, where $R_\beta = -r/\sin\beta$. Hence the special tanvolute σ^*_β is a concentric circle, shown in Figure 11.18b, of radius

$$R_* = \frac{r}{|\sin\beta|}. \tag{11.45}$$

General tanvolute of a circle.

Next we analyze (11.42) for large $|\alpha|$ when the exponential terms are present (which implies $\beta \neq 0$). Either the exponential term dominates (when $c\alpha > 0$), or it tends to 0 (when $c\alpha < 0$), in which case the linear term dominates. Geometrically, this means that a general tanvolute spirals away from the original circle in unbounded fashion in one direction, and tends asymptotically to the special circular tanvolute in the other direction. Figure 11.18 gives examples with $\beta = -30^o$, so $c < 0$.

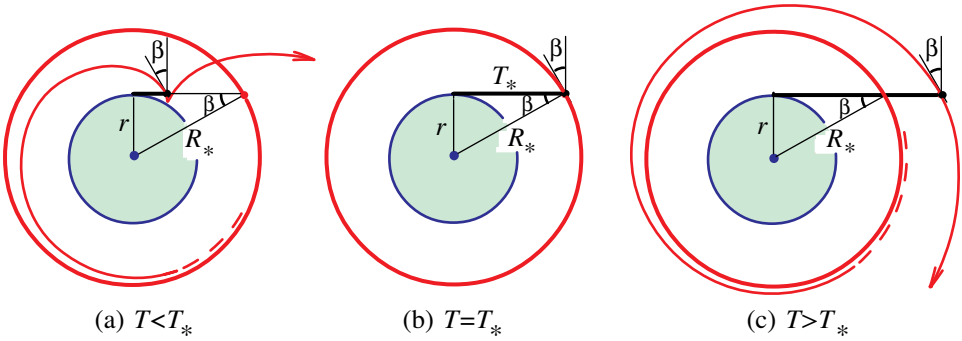


Figure 11.18: Here $\beta = -30°$. (a) When $T < T_*$, the tanvolute spirals outward, toward the special circular tanvolute if $c\alpha < 0$, but in unbounded fashion if $c\alpha > 0$. (b) When $T = T_*$, the tanvolute is the special tanvolute. (c) For $T > T_*$, the tanvolute spirals inward toward the special tanvolute if $\alpha > 0$, but spirals outward in unbounded fashion if $\alpha < 0$.

Example 2 (Tanvolutes of a single point are logarithmic spirals). Now we take the tangency curve to be a single point O, which can be regarded as the limiting case of a small circular arc τ whose radius shrinks to zero. As $r \to 0$ in (11.42) we find that the arclength function of the tanvolute is

$$s_\beta(\alpha) = \frac{T}{\sin\beta}(e^{c\alpha} - 1). \tag{11.46}$$

The coefficient $T/\sin\beta$ in (11.46) has a simple geometric meaning. Let L denote the arclength of the tanvolute between O ($\alpha = -\infty$) and the point on the tanvolute where $\alpha = 0$. From (11.46) we find $L = s_\beta(0) - s_\beta(-\infty) = T/\sin\beta$, and hence

$$s_\beta(\alpha) = L(e^{c\alpha} - 1).$$

Similarly, (11.43) gives the limiting case $t_\beta(\alpha) = Te^{c\alpha} = L\sin\beta\, e^{c\alpha}$, where $t_\beta(\alpha)$ becomes the radial distance $r_\beta(\alpha)$ from point O to σ_β. In general, the tanvolute has polar equation

$$r_\beta(\alpha) = Te^{c\alpha},$$

where $T = L\sin\beta$. When $\beta = 0$, $s_0(\alpha) = T\alpha$, giving a circle of radius T, as in Figure 11.19a. But when $\beta \neq 0$, this is the polar equation of a logarithmic spiral, shown in Figure 11.19b for $\beta > 0$. This agrees with the geometric definition of a logarithmic spiral as a curve cutting all its polar radial lines at a constant angle of attack.

Figure 11.19c shows the region between the spiral arc OA of length L and the initial segment OA, dissected into tiny triangles with equal vertex angles (as Archimedes did with a circular disk) and unfolded to fill part of right triangle AOB. This shows geometrically why AB has length L and why the region between the spiral arc OA and the initial tangent has area equal to half that of triangle OAB, a known result that can also be verified analytically.

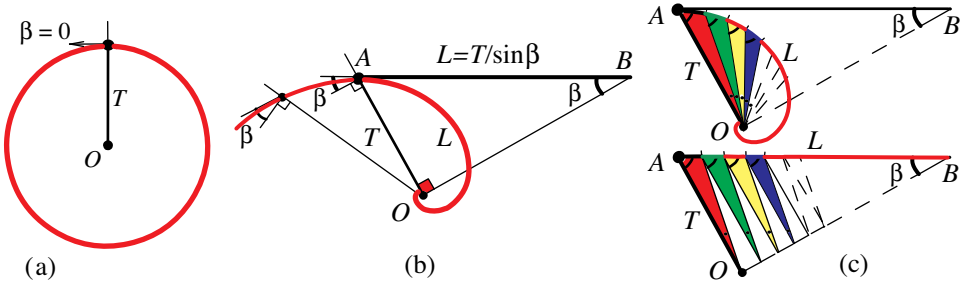


Figure 11.19: Tanvolutes of a single point: (a) Circle as involute. (b) Logarithmic spiral as a β-tanvolute. (c) Arclength and area obtained geometrically by an unfolding process.

Further examples illustrating Problem 1 are described in Section 11.17. They include tanvolutes of a logarithmic spiral, tanvolutes of the involute of a circle, tanvolutes of cycloids, epicycloids, and hypocycloids. Discussion is also given there of the relation between the initial tangent T and the resulting shapes of the tanvolutes.

11.13 TANVOLUTES APPLIED TO PURSUIT PROBLEMS

We turn now to an application generalizing a remarkable pursuit problem that appears in the literature in various forms and under various names, e.g., the trawler problem, the rum-runner problem, or the anti-missile problem (Exercise 16 in [1; p. 544]). The problem is of special interest because at first glance it seems that the given information is insufficient to solve it. We shall generalize all forms as an anti-missile problem, and solve it with the help of tanvolutes.

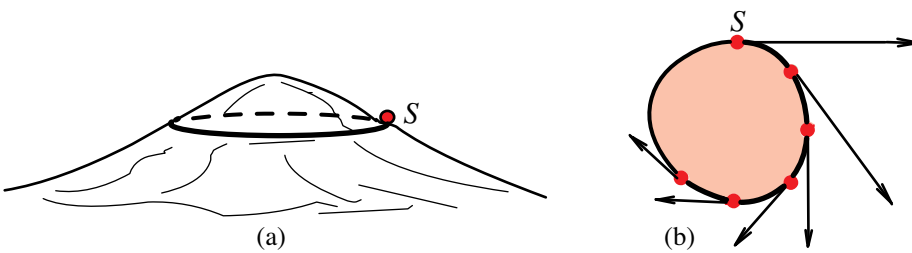


Figure 11.20: (a) Horizontal missile track on a mountain. (b) Top view showing the track as a convex tangency curve. The tangent vectors indicate possible directions of missile 1.

Generalized pursuit problem.

A military base is equipped with a horizontal convex track around a mountain (Figure 11.20a), along which a missile can fly (clockwise) at constant speed v and can be aimed, at the same speed v, toward a target along a line tangent to the track (Figure 11.20b). A missile, mistakenly fired, starts at point S and flies along the

track to an unknown point on the track from which it proceeds along the tangent direction, which is also unknown (Figure 11.20b). To destroy the first missile, an anti-missile missile is fired from S exactly two seconds later at constant speed $V > v$. Unlike the first missile, the second missile can fly along any path.

What path should the second missile follow to overtake the first one?

First, we try to determine where missile 1 could be at any moment. In Figure 11.21a, suppose that, at a particular moment, missile 1 has moved a distance a along the base track plus a distance b along the unknown direction of tangency. At that moment, missile 1 lies somewhere on the involute to the tangency track along which $a + b$ is constant, as though a string of length $a + b$ is unwrapped from the track. The dashed curves in Figure 11.21a show how the involutes expand uniformly with time.

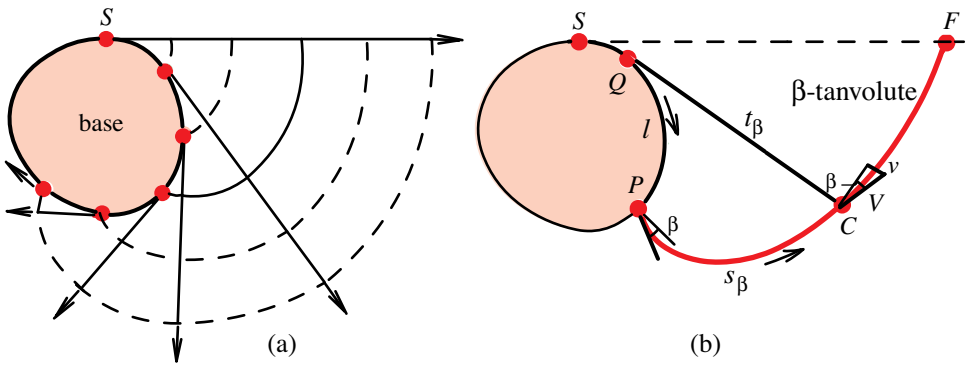


Figure 11.21: (a) The possible time-loci of missile 1 are involutes of the base track, expanding uniformly with time. (b) Missile 2 proceeds to point P on the base track where interception would occur if missile 1 was still on the track, and then follows the β-tanvolute of the track.

There are several strategies available to overtake missile 1. The simplest assumes that missile 1 travels from the starting point S and moves a distance SQ along the track to an unknown point Q, then proceeds along the tangent at Q, as in Figure 11.21b. Then missile 2 travels along the track from S to the point P where missile 1 would have been had it stayed on the track. Point P can be determined because the missile speeds v and V are known.

If missile 2 hits missile 1 at P, the pursuit is over. Otherwise, missile 2 changes course and moves from P along a β-tanvolute, which will intersect all the tangent lines along which missile 1 might travel. Of course, we need to choose β so that missile 2 collides with missile 1 at some point, denoted by C in Figure 11.21b. The small triangle suggests that β should be chosen so that $\sin\beta = v/V$. The reason is that missile 1 moves tangentially a distance v in unit time while missile 2 moves along the β-tanvolute a distance V, forming a leg and hypotenuse of a small right triangle with adjacent angle whose sine is v/V. This is the angle of attack β.

We can also obtain this choice of β analytically. Missiles 1 and 2 are to reach C

simultaneously. Missile 1 travels the distance SQ along the track plus the distance $t_\beta(\alpha)$ along the tangent ray. The time required to cover the distance at constant speed v is

$$\frac{SQ + t_\beta(\alpha)}{v}.$$

Now, missile 2 travels the distance SP from S to P, then the distance $s_\beta(\alpha)$ from P to C, where α is measured so that $\alpha = 0$ at P. The time required to cover the distance at constant speed V is

$$\frac{SP + s_\beta(\alpha)}{V}.$$

Because missile 2 is fired later than missile 1, say the time delay is Δ, collision will occur at C if

$$\Delta + \frac{SP + s_\beta(\alpha)}{V} = \frac{SQ + t_\beta(\alpha)}{v}. \tag{11.47}$$

But Δ is the difference in times required for the two missiles to travel from S to P:

$$\Delta = \frac{SP}{v} - \frac{SP}{V}.$$

Use this in (11.47), together with $SQ = SP - l(\alpha)$. Then (11.47) becomes

$$\frac{s_\beta(\alpha)}{V} = \frac{t_\beta(\alpha) - l(\alpha)}{v}.$$

Comparing this with (11.36), with $T = 0$, we find

$$\frac{v}{V} = \frac{c}{\gamma} = \sin\beta,$$

the same result suggested geometrically.

The pursuit problem can be solved similarly if missile 2 starts at a point off the track (now the time delay is not necessary). Missile 2 proceeds directly to the track and then follows the track to the point where missile 2 would hit missile 1 had missile 1 stayed on the track. If it hits, the pursuit is over. Otherwise, missile 2 follows the β-tanvolute at that point with $\sin\beta = v/V$. Collision will occur at the latest at F in Figure 11.21b.

In the limiting case when the tangency curve τ reduces to a single point, the expanding involutes are concentric circles (see Figure 11.19a), and the problem reduces to the classical anti-missile problem, or to the other two classical problems mentioned earlier. In each, the β-tanvolute is a logarithmic spiral, the tanvolute of a single point.

11.14 PROBLEM 2. FINDING TANGENCY CURVE WITH KNOWN β-TANVOLUTE

In Section 11.11, we specified the function l in differential equations (11.32) and (11.33) and solved for s and t, keeping β constant. Now we specify s, and determine

t and l, which we denote by t_β and l_β. In other words, Problem 2 specifies the β-tanvolute σ_β, because it prescribes s and β, and asks for the tangency curve τ (through its intrinsic equation $l = l(\alpha)$), as well as the corresponding tangent length function t. The geometric meaning of Problem 2 is illustrated in Figure 11.12b.

Problem 2. *For constant β and a function s with $s(0) = 0$, determine t_β and l_β satisfying* (11.32) *and* (11.33) *with initial condition*

$$l_\beta(0) = 0.$$

Solution. We obtain t_β at once from (11.33), which gives

$$t_\beta = \cos\beta \, \frac{ds}{d\alpha}.$$

Using this in (11.32) and integrating over the interval $[0, \alpha]$ we find

$$l_\beta(\alpha) = t_\beta(\alpha) - t_\beta(0) - \sin\beta \, s(\alpha).$$

Expressed in terms of s, this becomes

$$l_\beta(\alpha) = \cos\beta(s'(\alpha) - s'(0)) - \sin\beta \, s(\alpha).$$

When $\beta = 0$ this gives $l_0(\alpha) = s'(\alpha) - s'(0)$, and the last equation becomes

$$l_\beta(\alpha) = -s(\alpha)\sin\beta + l_0(\alpha)\cos\beta. \tag{11.48}$$

Thus, the arclength function l_β is a linear combination of the arclength functions for σ and the curve whose involute is σ, which is known as the evolute of σ. Equation (11.48) resembles the formula for rotating cartesian coordinate axes, except now the intrinsic coordinates $l_\beta(\alpha)$ of the tangency curve τ are obtained from the intrinsic coordinates of the involute and the tangency curve by a combination of rotation and scaling.

Similarly, we find $t_\beta(\alpha) = t_0(\alpha)\cos\beta$, which tells us that the length of the tangent segment from the τ to the β-tanvolute is smaller than the tangent segment from τ to the involute by a factor $\cos\beta$. The maximum tangent length occurs when $\beta = 0$. Of course, each example illustrating Problem 1 also provides an example illustrating Problem 2.

Evolutoids.

The tangency curve τ is the envelope of inclined normals to the tanvolute σ_β, the constant angle of inclination being β. The problem of finding the envelope of normals turned by a fixed angle to a given curve σ appeared in the literature in work by Réaumur in 1709 and by Lancret in 1811. (See [54], p. 297.) In this earlier work, the envelope is referred to as the *evolutoid* or *developpoid* of σ. For $\beta = 0$ the evolutoid becomes the classical evolute of σ.

Réaumur was the first to show that the evolutoid of a circle is a concentric circle. But apparently no one realized that many other curves have a given circle as

evolutoid. This is illustrated in Figure 11.18, which shows a family of β-tanvolutes of a given circle, each tanvolute obtained by varying the initial tangent length T but keeping β fixed. The circle is the evolutoid of each of these tanvolutes.

In the same way, for a tangency curve τ and any β, there is an infinite family of β-tanvolutes, obtained by varying the initial tangent length T. All members of the family have the same tangency curve τ as β-evolutoid but with different values of T. The simplest tanvolute is the special β-tanvolute, for which $T = T_*$, so that there is no exponential influence.

Figures 11.22a and 11.22b show the same cardioid (with interior shading) as evolutoid of two different β-tanvolutes for the same β, the second being the special β-tanvolute. In Figures 11.22b, c, and d, each of the smaller cardioids is an evolutoid of the same larger cardioid for different values of β.

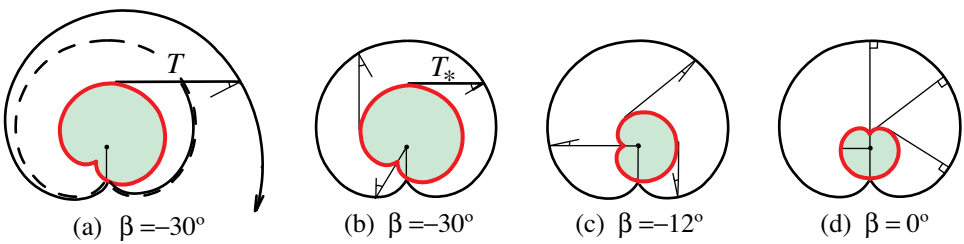


Figure 11.22: (a) General β-tanvolute of a cardioid. (b) Special tanvolute of the same cardioid for the same β. In (b), (c) and (d): The large cardioid is the special β-tanvolute of each smaller cardioid for three different values of β.

11.15 PROBLEM 3. FINDING β-TANVOLUTES WHEN t IS KNOWN

We turn next to the third in our trilogy of problems. For fixed β, this one specifies t and determines both s and l, which we denote by s_β and l_β.

Problem 3. *For constant β and a given function t with $t(0) = T$, determine s_β and l_β satisfying* (11.32) *and* (11.33) *with initial conditions*

$$s_\beta(0) = 0, \quad l_\beta(0) = 0.$$

Solution. By integrating (11.32) and (11.33), using $\gamma = 1/\cos\beta$ and $c = \tan\beta$, we find

$$s_\beta(\alpha) = \gamma \int_0^\alpha t(\theta)d\theta, \text{ and } l_\beta(\alpha) = t(\alpha) - T - c\int_0^\alpha t(\theta)d\theta.$$

This simple solution reveals a surprising geometric fact: A knowledge of the tangent-length function t alone determines intrinsically both the tangency curve and its β-tanvolute, except for their position and orientation in the plane. All examples illustrating Problems 1 and 2 also illustrate Problem 3.

In particular, if t is constant, say $t = T$, we find $s_\beta(\alpha) = \gamma T\alpha$ and $l_\beta(\alpha) = -cT\alpha$, each of which represents the arclength of a circle. If $t(\alpha)$ is a linear function of α,

say $t(\alpha) = T + r\alpha$, where T and r are constants, then $s_\beta(\alpha) = \gamma T\alpha + \gamma r\alpha^2/2$, and $l_\beta(\alpha) = r\alpha - cr\alpha^2/2$, each of which is the arclength of the involute of a circle.

11.16 GEOMETRIC BEHAVIOR OF β-TANVOLUTES

Now we return to constant β and analyze the geometric behavior of β-tanvolutes as revealed by the explicit formulas for the functions $s_\beta(\alpha)$ and $t_\beta(\alpha)$ in (11.40) and (11.41). We begin by pointing out some differences between tanvolutes with $\beta \neq 0$ and classical involutes with $\beta = 0$.

The exponential influence.

The exponential influence $e^{c\alpha}$ in these formulas is not present in the classical case $\beta = c = 0$. In the nonclassical case, $\beta \neq 0$, the exponential $e^{c\alpha}$ appears implicitly in the integral $I(\alpha)$ in (11.39), and also explicitly with a factor T in (11.40) and (11.41). As might be expected, this has a profound effect on the shape of nonclassical tanvolutes.

Canonical form.

In all examples treated in this paper, formula (11.40) for the arclength function $s_\beta(\alpha)$ can be written so that the exponential factor $e^{c\alpha}$ appears explicitly:

$$s_\beta(\alpha) = S_\beta(0)e^{c\alpha} - S_\beta(\alpha). \tag{11.49}$$

In the second term, $S_\beta(\alpha)$ is a function of c and α that does not involve $e^{c\alpha}$ explicitly. And its initial value $S_\beta(0)$ is the coefficient of $e^{c\alpha}$ in the first term! We call (11.49) the *canonical form* of (11.40). Both $S_\beta(\alpha)$ and $S_\beta(0)$ depend on the initial value T.

By differentiating (11.49) and using (11.33) we find

$$t_\beta(\alpha) = (\sin\beta)S_\beta(0)e^{c\alpha} - (\cos\beta)\frac{dS_\beta(\alpha)}{d\alpha}, \tag{11.50}$$

which we call the *canonical form* of (11.41).

Special tanvolutes with no exponential influence.

If $S_\beta(0) = 0$ for some choice of T, there is no exponential term in (11.49), in which case (11.49) becomes $s_\beta(\alpha) = S^*_\beta(\alpha)$, where $S^*_\beta(\alpha)$ is the function $-S_\beta(\alpha)$ with $S_\beta(0) = 0$.

We shall refer to the tanvolute having arclength function S^*_β as a *special tanvolute*, and we denote it by σ^*_β, as was done earlier in Example 1 in treating tanvolutes of a circle.

The canonical form (11.49) reveals the surprising fact that the arclength function $s_\beta(\alpha)$ of a general tanvolute is a linear combination of an exponential term $e^{c\alpha}$ and the arclength function of a special tanvolute.

For example, the arclength of the tanvolute of a circle as given in (11.42) can be written in canonical form (11.49) with

$$S_\beta(\alpha) = \frac{\gamma}{c}(r\alpha + T + \frac{r}{c}), \tag{11.51}$$

which represents the arclength function of a circle of radius $R_* = |\gamma r/c| = r/|\sin\beta|$.

In all our examples, a general tanvolute spirals away from the tangency curve in an unbounded fashion in one direction, and tends asymptotically to the special tanvolute in the other direction. The special tanvolute, in turn, is a scaled and rotated version of the tangency curve.

Attached and detached tanvolutes.

If $t_\beta(\alpha) = 0$ for some α we call the tanvolutes *attached* because the free end of the string is actually on the tangency curve τ for that value of α. Examples for a circle are shown in Figure 11.17 and in Figure 11.18a.

In Figures 11.17a, 11.17b, and 11.18a, the tanvolute has a cusp at the point of contact with the circle. This happens with an attached tanvolute when $\beta \neq \pm\pi/2$. From (11.33) we see that $t_\beta(\alpha) = 0$ implies $ds_\beta/d\alpha = 0$ when $\beta \neq 0$. We note that $ds_\beta/d\alpha$ is the radius of curvature of the tanvolute σ_β, and a zero radius of curvature produces a cusp.

If $t_\beta(\alpha) \neq 0$ for all α the tanvolute never touches τ and we call the tanvolute *detached*. Examples for a circle are shown in Figures 11.18b and c.

As we will see in many examples, detached tanvolutes exhibit special behavior that does not occur with attached tanvolutes.

The role of the initial value T.

For every β, our initial conditions require that $s_\beta(0) = 0$ and $t_\beta(0) = T$. The initial value T, which can be positive, negative, or zero, can be regarded as a free parameter in formulas (11.40) and (11.41). Geometrically, $|T|$ represents the tail of the string being unwrapped from τ. In classical treatments, conic sections are examples of detached involutes with a tail, whereas cycloidal involutes are attached with no tail. (See Figure 11.29a.) Treating T as a free parameter leads to more flexibility in several ways. For example, in the case of attached tanvolutes, changing T is equivalent to unwrapping the string from different points of tangency curve τ.

Also, for a given β, the tanvolute σ_β depends on the initial value T, and we indicate this dependence by writing $\sigma_\beta = \sigma_\beta(T)$. When the tanvolute is a spiral, different choices of T can produce the same spiral. In fact, the line through the initial tangent point intersects the spiral infinitely often at discrete points $T_k = t_\beta(2\pi k)$, where k is any integer, positive, negative, or zero. The same spiral σ_β is traced as tanvolute if we choose any T_k as initial value. This is equivalent to replacing α by $\alpha' = \alpha + 2\pi k$ in (11.40) and (11.41). For example, if $k > 0$ the portion of the spiral beyond T_k is obtained by allowing α' to increase (counterclockwise) from 0 to ∞. The earlier portion of the spiral is traced by allowing α' to decrease (clockwise)

from 0 to $-\infty$. This means that the same spiral σ_β is traced for each initial choice T_k. In symbols, $\sigma_\beta(T_0) = \sigma_\beta(T_k)$.

Now start with a spiral tanvolute $\sigma_\beta(T_0)$ and choose a new initial value T in any of the intervals $T_k < T < T_{k+1}$. As T varies through the interval, we obtain a family of intermediate spiral tanvolutes $\sigma_\beta(T)$. To avoid excessively large values of T we usually choose T in the interval $T_0 < T < T_1$ or in the interval $T_{-1} < T < T_0$. As T increases we obtain the special tanvolute σ_β^* when T reaches a certain value T_*, and as T increases further we might reach a critical value T_{cr} that marks the transition between attached and detached tanvolutes. In the next section we will see that these values coincide, $(T_* = T_{\text{cr}})$, when τ is a circle or a logarithmic spiral, as shown in Figures 11.19 and 11.23. Later we will see, in Figure 11.29 for example, that some tangency curves do not have detached tanvolutes.

11.17 FURTHER EXAMPLES ILLUSTRATING PROBLEM 1

Example 3 (Tanvolutes of a logarithmic spiral). Start with a logarithmic spiral with polar equation $r = Te^{a\alpha}$, where a is a positive constant given by $a = \tan\delta$, and δ is the complement of the constant angle between the tangent to the spiral and the radial line from the origin, so that δ plays the same role as β in Figure 11.19b. (For example, in Figure 11.13, $\delta = \pi/n$.) Its arclength $l(\alpha)$ has the form $l(\alpha) = L(e^{a\alpha} - 1)$, where the constant L is the arclength of the portion of the spiral from the origin to the point on the spiral where $\alpha = 0$. This implies $l(0) = 0$. Using this in the integrand for $I(\alpha)$ in (11.39) we find

$$I(\alpha) = \int_0^\alpha L(e^{a\theta} - 1)e^{c(\alpha-\theta)}d\theta = Le^{c\alpha}\int_0^\alpha e^{(a-c)\theta}d\theta + L\frac{1-e^{c\alpha}}{c}.$$

If $c = a$ this becomes

$$I(\alpha) = L\alpha e^{a\alpha} + L\frac{1-e^{a\alpha}}{a},$$

but if $c \neq a$ we have

$$I(\alpha) = L\frac{e^{a\alpha} - e^{c\alpha}}{a-c} + L\frac{1-e^{c\alpha}}{c}.$$

Using these in (11.40) we find the following formulas for the arclength of the tanvolute of a logarithmic spiral:

If $c = a$, then

$$s_\beta(\alpha) = \gamma(T-L)\frac{e^{a\alpha}-1}{a} + \gamma L\alpha e^{a\alpha}, \tag{11.52}$$

but if $c \neq a$, then

$$s_\beta(\alpha) = \gamma(T-L)\frac{e^{c\alpha}-1}{c} + \gamma L\frac{e^{a\alpha}-e^{c\alpha}}{a-c}. \tag{11.53}$$

These formulas are based on unwrapping the tanvolute from the logarithmic spiral in the counterclockwise direction, indicated by allowing α to increase from 0 to ∞. We could just as well unwrap the tanvolute in the opposite direction, indicated by allowing α to decrease from 0 to $-\infty$. From (11.53), the special tanvolute corresponds to $T = La/(a-c)$ with $S_\beta^*(\alpha) = \gamma L(e^{a\alpha} - 1)/(a-c)$. This is the original logarithmic spiral scaled by $\gamma/(a-c)$ and rotated by $\pi/2 + \beta$.

Involutes of a logarithmic spiral.

Now let $\beta \to 0$ in (11.53) and we get

$$s_0(\alpha) = (T - L)\alpha + L\frac{e^{a\alpha} - 1}{a} \tag{11.54}$$

as the arclength function of the involute. When $T = L$ the unwrapping takes place from the origin and the arclength function reduces to $s_0(\alpha) = L(e^{a\alpha} - 1)/a$, which is a scaled version of the arclength function of the original spiral. As $\alpha \to +\infty$ the involute spirals outward, but as $\alpha \to -\infty$ it spirals inward toward an asymptotic circle of radius $R = |L - T|$. Examples are shown in Figure 11.23.

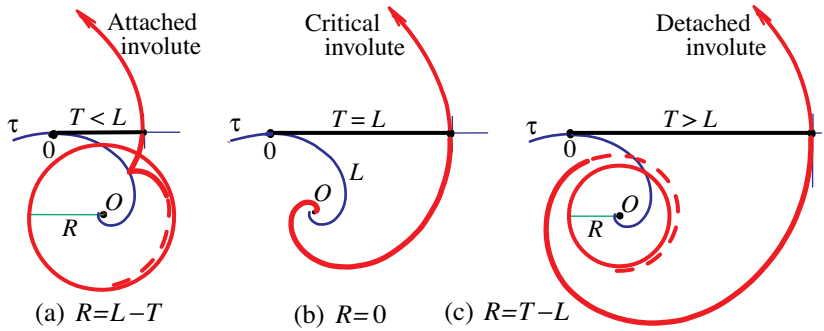


Figure 11.23: Involutes of a logarithmic spiral for (a) $T < L$, (b) $T = L$, (c) $T > L$.

Tanvolutes of a logarithmic spiral.

Let $E = T - La/(a - c)$. Figure 11.24 shows various attached tanvolutes of a logarithmic spiral with $\beta \neq 0$ and $T = 0$. The asymptotic behavior is obtained by analyzing (11.52) and (11.53) when $c = a$ and $c \neq a$, respectively.

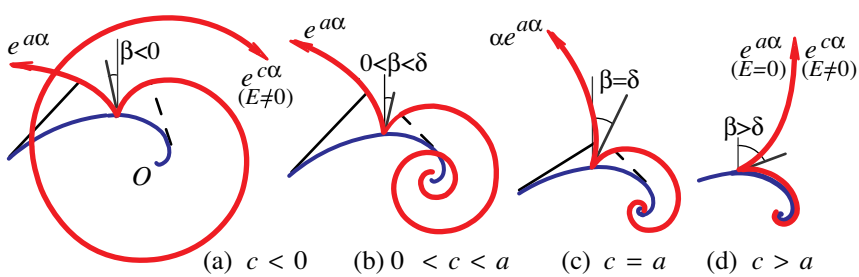


Figure 11.24: Attached β-tanvolutes of a logarithmic spiral for various values of β. Expanding asymptotic behavior is marked on each graph.

Figure 11.25 shows tanvolutes for a logarithmic spiral with $\beta < 0$ and $T > 0$. The asymptotic behavior depends on the relation between T and L, and can be analyzed by using (11.52) and (11.53). We omit the details.

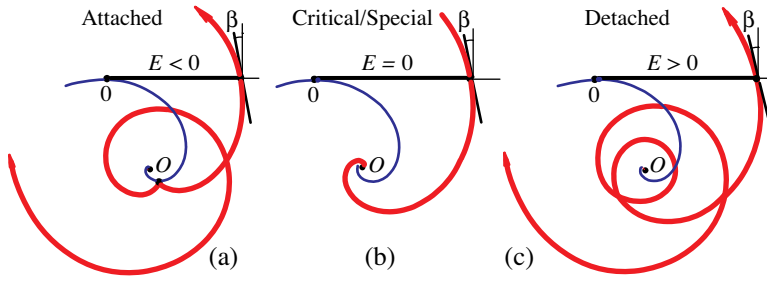


Figure 11.25: Tanvolutes of a logarithmic spiral with $\beta < 0$, $T > 0$. In (a) $E < 0$, in (b) $E = 0$, and in (c) $E > 0$.

Example 4 (Tanvolutes of the involute of a circle). In Example 1 we found the intrinsic equation of the involute of a circle of radius r to be

$$s_0(\alpha) = T\alpha + \frac{1}{2}r\alpha^2. \tag{11.55}$$

If $T = 0$ this becomes an even function of α, and in that case the involute σ is a symmetric curve with two branches meeting at a cusp at point O on the circle, which corresponds to $\alpha = 0$.

Now we use the involute of a circle as the tangency curve, with arclength function $l(\alpha)$ given by the right member of (11.55) with $T = 0$, and determine its tanvolutes. In this case the convolution integral in (11.39) is given by

$$I(\alpha) = e^{c\alpha}\int_0^\alpha \frac{1}{2}r\theta^2 e^{-c\theta}d\theta = \frac{r}{2}(-\frac{\alpha^2}{c} - \frac{2\alpha}{c^2} + \frac{2(e^{c\alpha}-1)}{c^3}),$$

which, when used in (11.40) with a general T, gives us

$$s_\beta(\alpha) = \gamma(T + \frac{r}{c^2})\frac{e^{c\alpha}-1}{c} - \gamma r(\frac{\alpha^2}{2c} + \frac{\alpha}{c^2}). \tag{11.56}$$

Involutes of the involute of a circle.

If $\beta \to 0$ then $c \to 0$, $\gamma \to 1$, and the limiting case of (11.56) (see Figure 11.26) gives the intrinsic equation for the involute of an involute of a circle: $s_0(\alpha) = T\alpha + \frac{r}{6}\alpha^3$.

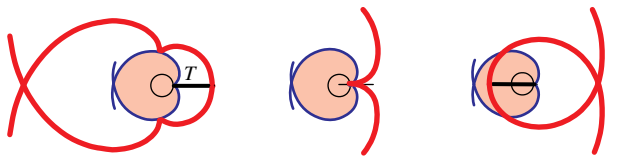


Figure 11.26: Classical case. Involutes of an involute of a circle.

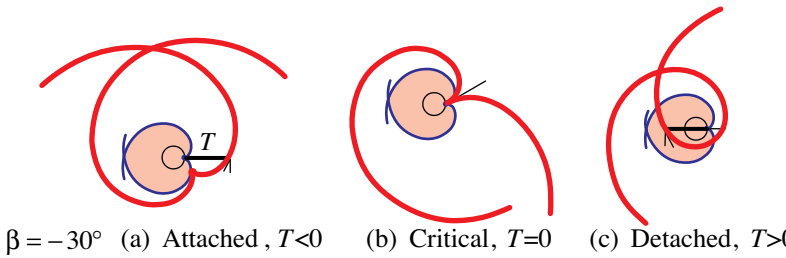


Figure 11.27: Tanvolutes of involute of a circle.

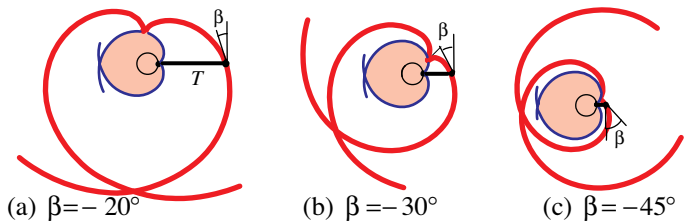


Figure 11.28: Special tanvolutes of involute of a circle.

Tanvolutes of the involute of a circle.

The examples in Figure 11.27 show how the tanvolute depends on the initial value T for a fixed $\beta < 0$.

Figure 11.28 shows tanvolutes with no exponential influence. To determine them, write (11.56) in the canonical form $s_\beta(\alpha) = S_\beta(0)e^{c\alpha} - S_\beta(\alpha)$, where

$$S_\beta(\alpha) = \frac{\gamma}{c}(T + \frac{r\alpha^2}{2} + \frac{r\alpha}{c} + \frac{r}{c^2}), \text{ and } S_\beta(0) = \frac{\gamma}{c}(T + \frac{r}{c^2}).$$

When $T = -r/c^2$ the exponential influence disappears and $s_\beta(\alpha)$ is a quadratic function of α. In this case, the special tanvolute of the involute of a circle of radius r is an involute of another circle of radius $R_* = r/|\sin\beta|$.

Example 5 (Tanvolutes of cycloids, epicycloids, and hypocycloids). Earlier we have shown that the arclength function $l(\alpha)$ for a cycloidal, epicycloidal, or hypocycloidal arc has the form

$$l(\alpha) = A\sin b\alpha, \tag{11.57}$$

where $A = 4r(1 \pm r/R)$, $b = 1/\kappa$, and $\kappa = 1 \pm 2r/R$. When a disk of radius r rolls outside a circle of radius R to produce the epicycloid we use the plus sign, and when it rolls inside to produce the hypocycloid we use the minus sign. The cycloid is the limiting case $R = \infty$. For the cycloid, $l(\alpha)$ is the length of the cycloidal arc from the highest point M of a cycloidal arch to a point P where the tangent segment makes an angle of inclination α with the horizontal line through M. There is a similar interpretation for $l(\alpha)$ for the epicycloid and hypocycloid.

The constant A, the amplitude of the sine function in (11.57), is equal to the length of one half of the cycloidal, epicycloidal, or hypocycloidal arch. Now insert the general arclength formula (11.57) in (11.39) and obtain

$$I(\alpha) = Ae^{c\alpha}\int_0^\alpha e^{-c\theta}\sin b\theta\ d\theta = \frac{A}{b^2+c^2}(be^{c\alpha} - b\cos b\alpha - c\sin b\alpha).$$

Formula (11.40) for the arclength function of the β-tanvolute now becomes

$$s_\beta(\alpha) = \frac{\gamma}{c}e^{c\alpha}E - \frac{\gamma}{c}T - \frac{\gamma A}{b^2+c^2}(b\cos b\alpha + c\sin b\alpha), \tag{11.58}$$

where $E = T + Abc/(b^2+c^2)$. Again, $E = S_\beta(0)\sin\beta$ from the canonical form (11.49). If $E = 0$ there is no exponential influence in (11.58). This occurs only if $T = -Abc/(b^2+c^2)$.

To get the classical involutes of these curves we let $\beta \to 0$. Then $c \to 0$ and $\gamma \to 1$, and (11.58) gives

$$s(\alpha) = T\alpha + \frac{A}{b}(1 - \cos b\alpha). \tag{11.59}$$

for the arclength of the involute of a cycloidal, epicycloidal, or hypocycloidal curve, a combination of the arclength of a circle and of a cycloidal curve of the same type.

Example 5a (Tanvolutes of a cycloid ($b = 1$)). Figures 11.29a-d are examples of attached involutes of a cycloid with various increasing values of the ratio T/A, which compares the initial tangent length T with A, half the length of a cycloidal arch.

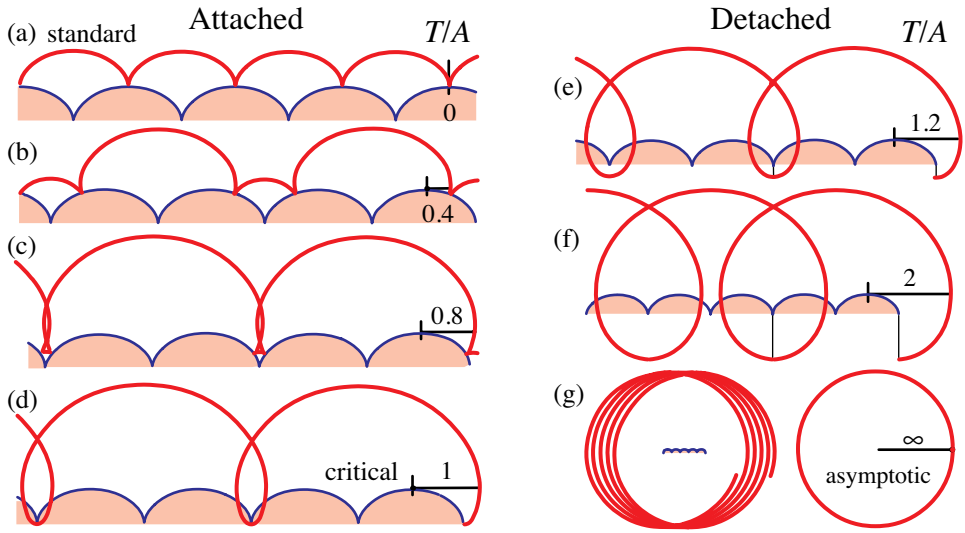


Figure 11.29: Involutes of a cycloid with various tangent length ratios T/A.

Figures 11.29e-g are examples of detached involutes of a cycloid. As the initial tangent length T increases, the linear term $T\alpha$ in (11.59) is dominant. It represents the arclength function of a circle. Apparently, only the case $T = 0$ in Figure 11.29a has been previously treated.

When $b = 1$, $b^2 + c^2 = 1 + c^2 = \gamma^2$, and the arclength function in (11.58) for the β-tanvolute of a cycloid becomes

$$s_\beta(\alpha) = \frac{\gamma}{c} e^{c\alpha} E - \frac{\gamma}{c} T - A\cos(\alpha - \beta), \tag{11.60}$$

where now the formula for E simplifies to $E = T + Ac/\gamma^2$. When $E = 0$, the exponential term disappears in (11.60) and $s_\beta(\alpha)$ is a cosine function that represents the tanvolute of a shifted cycloid. If $E \neq 0$ and $c\alpha > 0$, the exponential term dominates, but if $c\alpha < 0$ the shifted cycloidal term dominates. The tanvolute makes larger and larger spirals as $\alpha \to +\infty$, but stabilizes to a shifted cycloid as $\alpha \to -\infty$.

Figure 11.30 shows an example of a β-tanvolute of a cycloid with a small $|\beta|$, and E close to 0.

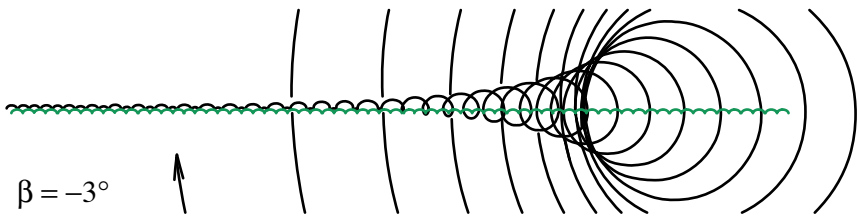


Figure 11.30: A β-tanvolute of a cycloid with small $|\beta|$ and E close to 0.

Figure 11.31 shows examples with larger values of $|\beta|$ and illustrates how the exponential influence of the tanvolute is pushed to ∞ as $E \to 0$, with the shifted cycloid being dominant. Each dominant shifted cycloid is a special tanvolute.

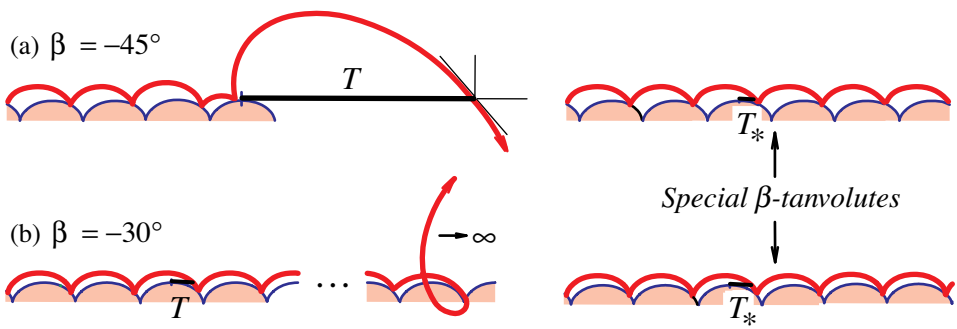


Figure 11.31: Two β-tanvolutes of a cycloid, with the shifted cycloid becoming more dominant as $E \to 0$. Special β-tanvolutes (no exponential influence) also shown.

Example 5b (Tanvolutes of an astroid ($b = 2$)). A hypocycloid with $r/R = 1/4$, $\kappa = 1/2$, $b = 2$, is called an astroid. Figure 11.32a shows attached involutes corresponding to increasing values of the ratio T/A from 0 to 1. Figure 11.32b shows detached involutes with various ratios T/A greater than 1. As $T \to \infty$ the involutes become more and more circular.

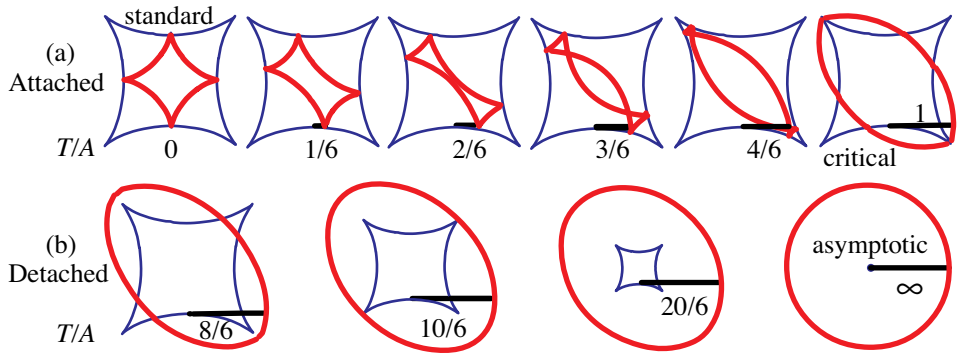


Figure 11.32: Involutes of an astroid with various initial tangent ratios T/A.

Figure 11.33 displays three β-tanvolutes with $\beta = -\pi/6$ and various values of the ratio T/A. In Figure 11.33b there is no exponential influence, and the tanvolute is a special tanvolute, which is a similar astroid, inscribed in the larger astroid, and rotated. Figure 11.34 shows a β-tanvolute of an astroid with $\beta = -\pi/60$, a value small enough so that some of the spirals appear on the page.

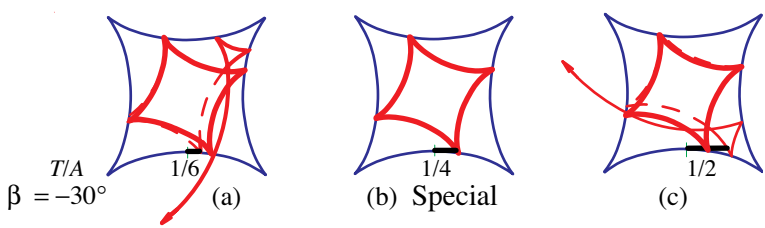


Figure 11.33: β-tanvolutes of astroid with $\beta = -\pi/6$ and various ratios T/A.

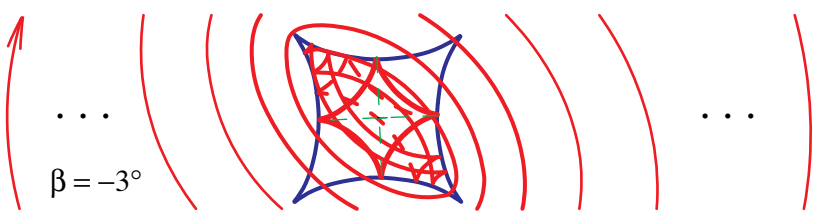


Figure 11.34: β-tanvolute of an astroid with a small $|\beta|$.

Example 5c (Tanvolutes of a cardioid ($b = 1/3$)). A cardioid is an epicycloid with $r = R$, $\kappa = 3$,and $b = 1/3$. Figure 11.35 shows some of its involutes corresponding to various initial tangent lengths. The leftmost involute in the top

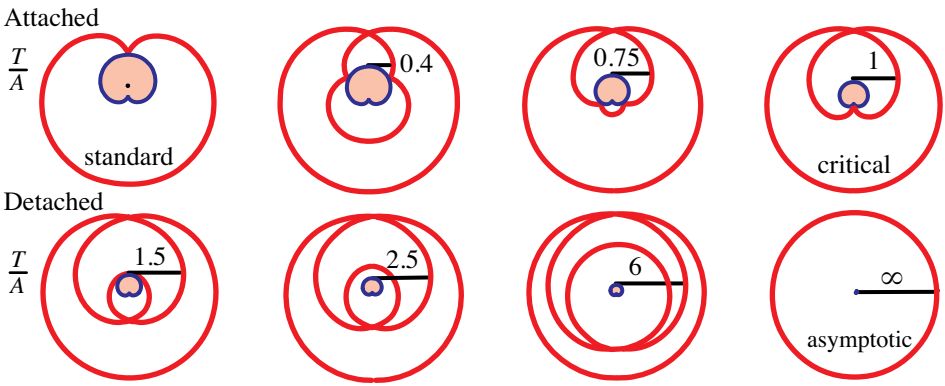


Figure 11.35: Involutes of a cardioid with different initial tangent lengths.

row of Figure 11.35 is the classical involute with initial length $T = 0$, which is usually the only involute treated. But there are more involutes possible when $T \neq 0$, some of which are shown. As T/A increases from 0 to 1, the involutes are attached, as shown by the examples in the first row. The second row shows examples of detached involutes obtained as T/A increases beyond 1.

Figure 11.36 shows tanvolutes of a cardioid (interior shaded) with various initial tangents. The special tanvolute in Figure 11.36b, which has no exponential influence, is a larger similar cardioid, rotated, with its cusp touching the smaller cardioid.

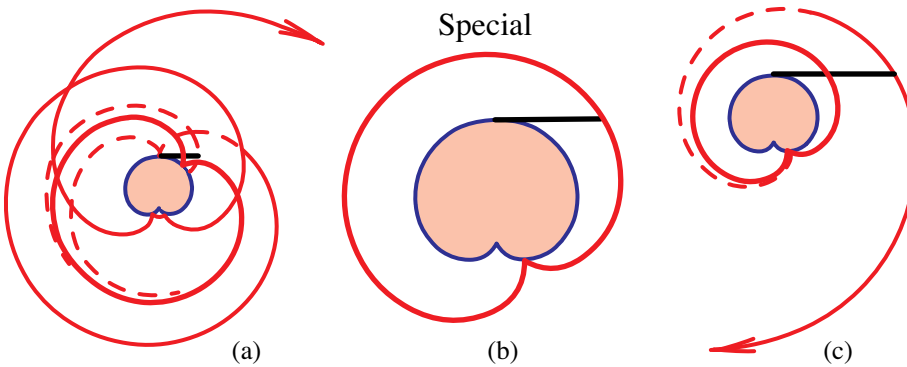


Figure 11.36: β-tanvolutes of a cardioid.

11.18 CUSPS OF CYCLOIDAL SPECIAL TANVOLUTES

Each special tanvolute of a cycloidal curve, which can be a cycloid, epicycloid, or hypocycloid, is a curve of the same type, similar and rotated. As β varies, the cusp of the β-tanvolute of a curve σ traces a path that can be determined explicitly. Figure 11.37 shows the locus of cusps of some examples.

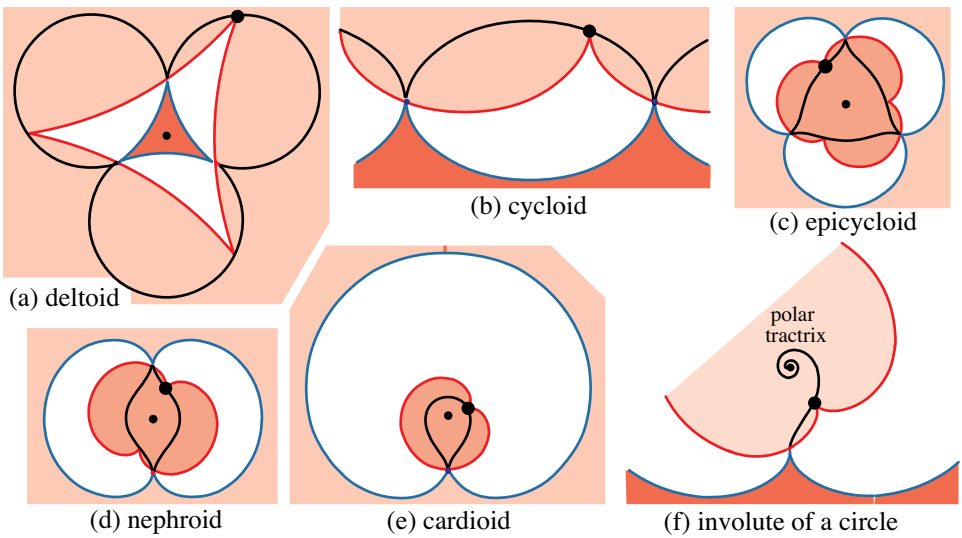


Figure 11.37: Locus of cusps of some cycloidal, epicycloidal, and hypocycloidal special tanvolutes.

In Figure 11.37a the locus of cusps of a deltoid consists of three arcs resembling circular arcs. In Figure 11.37b the locus is a reflected cycloid. In Figure 11.37c the locus is a curvilinear triangle, and in Figure 11.37d it consists of a symmetric loop. For the cardioid in Figure 11.37e the locus is a loop resembling an inverted teardrop. In Figure 11.37f the locus is a spiral known as a polar tractrix, a curve whose representation in polar coordinates has a constant polar tangent segment.

It is easy to show that the path traced by the cusp of these β-tanvolutes is the inversion of the curve σ through the circle circumscribing σ having its center at the center of the fixed circle defining the cycloidal curve (this circle being a straight line in the case of the cycloid). A proof can be given by referring to Figure 11.38.

In Figure 11.38a, the arc AV plays the role of a cycloidal arc with a cusp at A and a radial axis of symmetry through O, so that $OA = OV$. (The proof uses only this symmetry and not the fact that the arc is cycloidal.) In Figure 11.38b the arc is rotated to produce arc $A'V'$, and then in Figure 11.38c the rotated arc is scaled radially from O to become arc $A''V''$ where the scaled version passes through V. Point M is the point of intersection of AV and the radial line OA''. Therefore the curvilinear triangle OMV is similar to the curvilinear triangle $OA''V$ and we have $OA''/OV = OV/OM$, or $OA'' = (OV)^2/OM$. Now OV is the radius of the circle

with center O through A and V, hence the locus of A'' is the inversion of AMV with respect to the circle, as asserted.

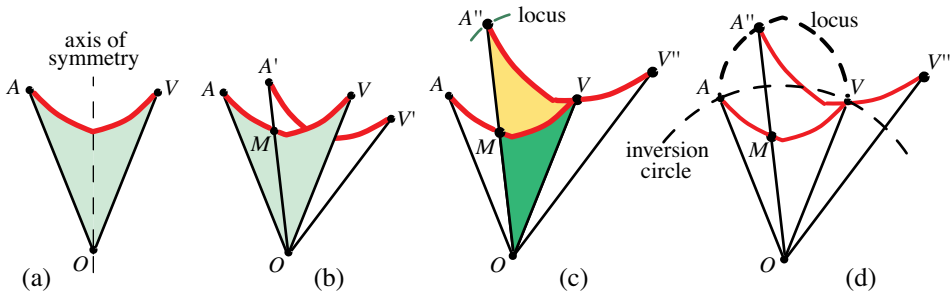


Figure 11.38: Proof that the locus of cusps is obtained by inversion.

11.19 VARIABLE ANGLE OF ATTACK β

Beginning in Section 11.10 we considered only a constant angle of attack β. But the basic differential equations (11.32) and (11.33) in Theorem 11.1 also hold when β is not constant. This opens the door to further research by choosing β to be a function of α.

In the general discussion it is important to establish the tangency angle ψ of the curve σ so that (ψ, s) provide intrinsic coordinates of σ. From a diagram similar to that in Figure 11.15 it is easy to see that in the most general case the tangency angle ψ (with respect to some initial direction) is given by

$$\psi = \alpha - \beta,$$

where α and β are as in Figure 11.15b, with β not necessarily constant.

In Sections 11.4 and 11.5 we treated several examples with $\beta = \alpha$, in which case $\psi = 0$ and the free-end curve σ is a straight line. This led to simple derivations of known formulas for the arclength $l = l(\alpha)$ of a cycloid (proportional to sin α), catenary (proportional to tan α), tractrix (proportional to log(cos α)). More complicated arclength functions were given for an exponential curve, a parabola, and generalized pursuit curves, all examples with $\beta = \alpha$.

Further examples treated in Section 11.5 are cycloidal curves with β proportional to α. The arclength of an epicycloidal arc traced by a point on the boundary of a disk of radius r rolling along the outer circumference of a fixed circle of radius R uses $\beta = \alpha/\kappa$, where $\kappa = 1+2r/R$. A corresponding result was obtained for the arclength of a hypocycloidal arc in which the circle of radius r rolls inside the circumference of the fixed circle of radius R. In this case $\beta = \alpha/\kappa$, where $\kappa = 1 - 2r/R$. In both cases, $\psi = \alpha(1 - 1/\kappa)$ and the arclength is proportional to sin β.

NOTES ON CHAPTER 11

Part 1 of this chapter (Sections 11.1 through 11.7) is based on work published by the authors in [22]. Most of the material in Part 2 (Sections 11.8 through 11.15 and 11.19) first appeared in [26]. The material in Sections 11.16 through 11.18 has not been previously published.

Part 1 uses sweeping tangents to obtain arclength and intrinsic equations of curves. The results in this chapter underscore the importance of intrinsic equations. Every plane curve can be regarded as the path of a moving particle. The intrinsic equation connects the distance covered by the particle with the angle through which the tangent line turns, and provides the most natural description of the curve because it does not rely on an external coordinate system. Our treatment of tanvolutes in Part 2 was simplified by using the intrinsic equations in Part 1.

When we classify the curves treated in this chapter according to their arclength functions $l = l(\alpha)$, we find a hierarchy of elementary functions. For a circle of radius r, $l(\alpha) = r\alpha$, a linear function of α. The arclength function for the involute of a circle is quadratic in α, and for the involute of the involute of a circle, it is cubic in α. Cycloids, epicycloids, and hypocycloids have sinusoidal arclength functions: $l(\alpha) = A \sin b\alpha$. A catenary has arclength function proportional to $\tan \alpha$, and the arclength function of a logarithmic spiral is an exponential of the form $l(\alpha) = L(e^{a\alpha} - 1)$.

In general, classical involutes are more complicated curves than their evolutes. For example, the arclength function of a circle in linear in α but that of its involute is quadratic in α. In fact, if the arclength function of a tangency curve τ is a polynomial in α of degree k, then that of its involute (obtained by integration) is a polynomial of degree $k + 1$. In contrast, the arclength function of the special tanvolute of τ involves a polynomial of degree k.

In all our examples, a general tanvolute spirals away from the tangency curve in an unbounded fashion in one direction, and tends asymptotically to the special tanvolute in the other direction. The special tanvolute, in turn, is a scaled and rotated version of the tangency curve. Particular examples are the special tanvolute of a circle, a concentric circle as in Figure 11.18b, and the special tanvolute of a cardioid, a scaled and rotated cardioid, as in Figure 11.22. In Figure 11.13, the special β-tanvolute of a logarithmic spiral is the same spiral rotated by $\beta + \pi/2$.

In the three main problems considered in Part 2 of this chapter, the tangent length function $t(\alpha)$ was either given or determined from the arclength functions l and s. A knowledge of $t(\alpha)$ allows us to find the area of the region between a base curve and its β-tanvolute swept by the tangent segment, as described in Chapter 1.

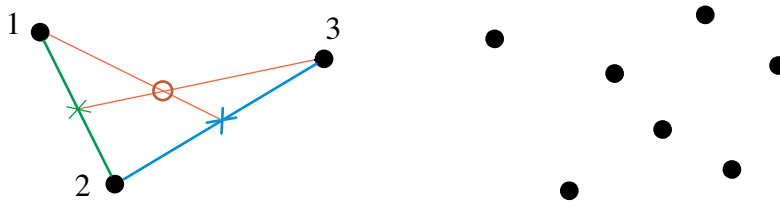
Finally, we note a relation between tanvolutes and bicycle wheels. The spokes of a bicycle wheel connecting the outer rim to a small axial disk are not positioned radially, but tangentially to the inner disk. Instead of forming right angles with the outer rim, the spokes are inclined at some small constant angle of attack, so the outer rim is like the special tanvolute of the inner disk. The technical reason for this arrangement of spokes is to keep them under tension rather than compression.

Chapter 12

CENTROIDS

These problems can be easily solved by the methods developed in this chapter. The reader may wish to try solving them before reading the chapter.

A lemma of Archimedes states that the centroid of a composite object consisting of two parts lies on the line segment joining the centroids of each of the two parts. The left part of the figure shows how to find the centroid of three points of equal mass using this lemma.



Choose two points, say 1 and 2; their centroid is midway between them. By the lemma of Archimedes, the centroid of all three points lies on the segment joining point 3 and the centroid of 1 and 2. Repeat the process starting with points 2 and 3. The centroid of all three points also lies on the segment joining point 1 and the centroid of 2 and 3. The two segments so constructed intersect at the centroid of all three points. Note that this method uses only bisection of segments and connecting points.

Determine the centroid of the seven points of equal mass in the right part of the figure by using only bisection of segments and connecting points.

Prove that the centroid of n points of equal mass can be found in a similar manner, using only $n-1$ bisections.

CONTENTS

PART 1: CENTROIDS OF PLANE FIGURES

PART 2: CENTROIDS OF n POINTS

The first part of the chapter shows how centroids of various plane figures can be easily determined without calculus. The second part describes two methods for locating the centroid of a finite number of points by graphical construction, without using coordinates or numerical calculations. The first method makes one arbitrary guess and forms a closed polygon; then an appropriate correction yields the exact location of the centroid. The second is an inductive procedure that combines centroids of two disjoint sets to determine the centroid of their union. It uses only bisection of line segments and connecting pairs of points. Both methods can be applied with drawing instruments or with computer graphic programs.

PART 1: CENTROIDS OF PLANE FIGURES

12.1 INTRODUCTION

Archimedes introduced the concept of center of gravity. He used it in many of his works, including the stability of floating bodies, ship design, and in his discovery that the volume of a sphere is two-thirds that of its circumscribing cylinder. It was also used by Pappus of Alexandria in the 3rd century AD in formulating his famous theorems for calculating volume and surface area of solids of revolution. We can only speculate on what Archimedes had in mind when he referred to center of gravity because none of his extant works provides an explicit definition of the concept. The existence of a center of gravity for every object is by no means obvious. Its discovery and use by Archimedes to solve problems in physics and mathematics is a landmark contribution to human thought.

Today a more general concept, center of mass, plays an important role in physics. Newton's laws of mechanics are usually stated so they apply to a point particle described by its position, velocity, and acceleration. Every object we see in nature is really a compound body made up of smaller parts, ultimately of atoms, and

physicists apply Newton's laws to such extended objects. They do this by imagining a large body, such as a planet or sun, as a single point where all its mass is concentrated. This point is called the center of mass. The position, velocity, and acceleration of a compound body are really those of the center of mass. In uniform centrally symmetric bodies the center of mass is identified with the center of symmetry.

We begin this chapter by defining the center of mass of a finite set of points. Given n points $\boldsymbol{r}_1,\ldots,\boldsymbol{r}_n$, regarded as position vectors in Euclidean m-space, relative to some origin O, let $w_1,\ldots,w_n$, be n positive numbers regarded as weights attached to them. The center of mass is the position vector $\boldsymbol{c}$ defined to be the weighted average given by

$$\boldsymbol{c} = \frac{1}{W_n}\sum_{k=1}^{n} w_k\boldsymbol{r}_k, \tag{12.1}$$

where W_n is the sum of the weights,

$$W_n = \sum_{k=1}^{n} w_k. \tag{12.2}$$

When all weights are equal, the center of mass $\boldsymbol{c}$ is called the centroid. If each $w_k = w$, then $W_n = nw$, the common factor w cancels in (12.1), and we get

$$\boldsymbol{c} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{r}_k, \tag{12.3}$$

which is equivalent to assigning weight 1 to each point. If the points $\boldsymbol{r}_1,\ldots,\boldsymbol{r}_n$, are specified by their coordinates, the coordinates of $\boldsymbol{c}$ can be obtained by equating components in (12.1).

Two important properties of center of mass follow at once from (12.1). Suppose the position vectors $\boldsymbol{r}_1,\ldots,\boldsymbol{r}_n$ are drawn from the origin O of a coordinate system. If the origin is translated to a new position by a vector $\boldsymbol{a}$, each position vector $\boldsymbol{r}_k$ is replaced by $\boldsymbol{r}_k - \boldsymbol{a}$ so the center of mass relative to the new origin is equal to

$$\frac{1}{W_n}\sum_{k=1}^{n} w_k(\boldsymbol{r}_k - \boldsymbol{a}) = \frac{1}{W_n}\sum_{k=1}^{n} w_k\boldsymbol{r}_k - \boldsymbol{a} = \boldsymbol{c} - \boldsymbol{a}.$$

In other words, the center of mass is shifted by the same amount as each position vector. This means that the center of mass is an intrinsic property of the body that does not depend on the origin of the coordinate system.

Next, if each position vector $\boldsymbol{r}_k$ is expanded or contracted by the same scaling factor t, then (12.1) shows that the center of mass $\boldsymbol{c}$ gets multiplied by the same factor t. We call this the *scaling property* of center of mass.

When we deal with a system whose total mass is distributed along all the points of an interval, or along a curve, or throughout some region in the plane or in space rather than at a finite number of discrete points, the concepts of mass and center

of mass are defined by integrals rather than sums. The definitions can be found in most calculus textbooks. For uniform bodies the center of mass is called the centroid.

Sections 12.2 through 12.4 determine centroids of several plane figures without using the formal machinery of integral calculus. Examples include the vertices of a triangle, an L-shaped lamina (a lamina can be thought of as a plane figure cut out of a thin piece of cardboard or plywood of uniform thickness), the boundary of a triangle, a triangular lamina, and a circular arc.

For a lamina with a center of symmetry, the centroid is at the center. To find the centroid of nonsymmetric laminas we shall use some basic properties of centroids, described below, together with a more profound property that we call the Archimedes Lemma, because it resembles Proposition 4 in [47; *Book I, On the Equilibrium of Planes*, p. 191].

12.2 THE ARCHIMEDES LEMMA

Archimedes Lemma. *If an object is divided into two smaller objects, the center of mass of the compound object lies on the line segment joining the centers of mass of the two smaller objects.*

For objects made up of a finite number of discrete masses a simple proof follows at once from the definition in (12.1), and the same type of proof works for more general mass distributions defined by integrals. We don't know how Archimedes proved Proposition 4 because he alludes to a previously proved property in his lost treatise *On Levers*.

Example 1 (Centroid of the vertices of a triangle). Figure 12.1a shows a system of three points, the vertices of a triangle. We will show that the centroid of this system is at the intersection of the three medians of the triangle. (A median is the line segment from a vertex to the midpoint of the opposite side.) First, consider a pair of vertices as one object, whose centroid is midway between the vertices, and the opposite vertex as the second object, and apply the Archimedes Lemma. This places the centroid $\boldsymbol{c}$ of the combined object on a median. Repeat the argument for each pair of vertices and we see that $\boldsymbol{c}$ is on all three medians. Therefore the three medians of a triangle intersect at a point, which is the centroid of the vertices. Later we will show that the centroid of the vertices is also the centroid of the entire triangular lamina.

To find the distance from a vertex to the centroid, refer to the parallelogram formed by the dotted lines in Figure 12.1b. Its upper edge joins two midpoints of the original triangle. Its lower edge joins midpoints of the inner triangle having two vertices at the original base and one vertex at the centroid. Two medians pass through the diagonals of this parallelogram. The diagonals bisect each other, so the intersection point of the two medians trisects both of them. In other words, the distance from each vertex to the centroid of the vertices is two-thirds the length of the median.

In general, the Archimedes Lemma involves three objects A, B, C whose centers

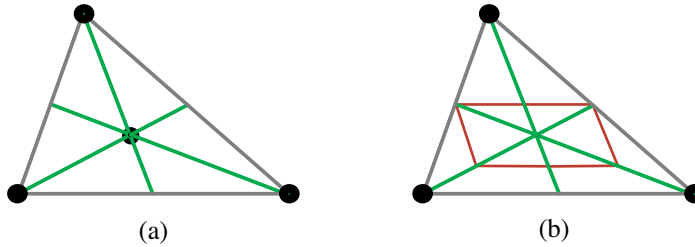


Figure 12.1: The centroid of the vertices of any triangle is at the intersection of the medians.

of mass $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$ lie on a line. But a line is uniquely determined by any two of its points. Therefore, if we know the center of mass $\boldsymbol{c}$ of a compound object C and the center of mass $\boldsymbol{b}$ of one of its parts B, then the center of mass $\boldsymbol{a}$ of the other part A is on the line joining $\boldsymbol{b}$ and $\boldsymbol{c}$. This variation of the Archimedes Lemma vastly increases the scope of its applicability, as we will demonstrate later with a number of examples involving plane laminas.

Besides the Archimedes Lemma we use several simple properties of centroids, each of which seems intuitively evident when the centroid of a plane lamina is regarded as the point about which the gravitational forces acting on the lamina are balanced.

12.3 FUNDAMENTAL PROPERTIES OF CENTROIDS OF PLANE LAMINAS

1. Translation property. Translation of a lamina by a vector translates its centroid to the centroid of the translated lamina.

2. Rotation property. Rotation of a lamina through an angle about an axis perpendicular to the plane of the lamina rotates its centroid to the centroid of the rotated lamina.

3. Centroid of a symmetric object. If a lamina has a center of symmetry, the centroid of the lamina is at the center of symmetry.

4. Centroid of a lamina with an axis of symmetry. If a lamina has an axis of symmetry, its centroid lies somewhere on it.

5. Centroid of an object formed from a lamina and any of its translates. A lamina together with any of its translates is a compound object whose centroid is at the midpoint of the line segment joining the centroids of the two laminas.

It is easy to verify these properties for objects consisting of a finite number of discrete points in a plane, using (12.1). Similar proofs can be given for the more general case in which centroids are defined by integrals. We will give proofs only

for properties (**2**) and (**5**) for objects consisting of a finite number of points.

Proof of (**2**). For this proof we represent each point $\boldsymbol{r}_k$ as a complex number. By property (**1**) there is no loss of generality if we assume the rotation axis passes through the origin, in which case rotation of $\boldsymbol{r}_k$ through an angle θ is the same as multiplying $\boldsymbol{r}_k$ by $e^{i\theta}$, so the centroid of the rotated system is equal to

$$\frac{\boldsymbol{r}_1 e^{i\theta} + \cdots + \boldsymbol{r}_n e^{i\theta}}{n} = \boldsymbol{c}\, e^{i\theta}.$$

Proof of (**5**). If the first object has points $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$ with centroid $\boldsymbol{c}$ then the translated object has points $\boldsymbol{r}_1 + \boldsymbol{a}, \ldots, \boldsymbol{r}_n + \boldsymbol{a}$, where $\boldsymbol{a}$ is the translation vector. The composite object consists of $2n$ points whose centroid $\boldsymbol{C}$ is given by

$$\boldsymbol{C} = \frac{\boldsymbol{r}_1 + \cdots + \boldsymbol{r}_n + (\boldsymbol{r}_1 + \boldsymbol{a}) + \cdots + (\boldsymbol{r}_n + \boldsymbol{a})}{2n}$$

$$= \frac{\boldsymbol{r}_1 + \cdots + \boldsymbol{r}_n}{n} + \frac{1}{2}\boldsymbol{a} = \boldsymbol{c} + \frac{1}{2}\boldsymbol{a},$$

so $\boldsymbol{C}$ is at the midpoint of the segment joining $\boldsymbol{c}$ and $\boldsymbol{c} + \boldsymbol{a}$.

12.4 APPLICATIONS

We have already used the Archimedes Lemma to find the centroid of the vertices of a triangle. We use it again to find the centroids of a variety of nonsymmetric objects. All the objects considered are assumed to be uniform bodies for which the center of mass and the centroid are identical. As mentioned above, the centroid of a uniform symmetric object, such as a rectangular lamina or a parallelogram, is at the center of symmetry.

Example 2 (L-shaped lamina). Figures 12.2a and 12.2b show an L-shaped plane lamina decomposed in two different ways into rectangular pieces. In each case, the centroid of the lamina lies on the line segment joining the centroids of the two rectangular pieces. Consequently, the centroid of L is at the point of intersection of these line segments, as shown in Figure 12.2c.

In practice there may be some difficulty in determining the point of intersection of the two line segments in Figures 12.2a and 12.2b because they could be nearly parallel. Figure 12.3 shows another way to find it. Extend the L-shaped lamina as shown to form a large rectangle whose centroid is at its center.

The centroid of the smaller rectangle in the corner is at *its* center. Therefore, by the variation of the Archimedes Lemma, the centroid of L must also lie on the line joining these two centers. This line intersects each of the lines in Figures 12.2a and 12.2b at the required centroid of L.

Example 3 (Two intersecting line segments, a limiting case of Example 2). When each rectangle of the L-shaped lamina in Example 2 degenerates to a line segment, the object consists of two perpendicular line segments. We can easily find the centroid of such an object even if the segments are not perpendicular.

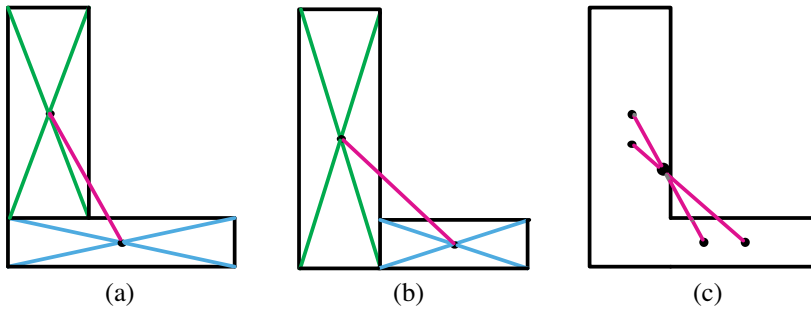


Figure 12.2: Centroid of an L-shaped lamina found by using the Archimedes Lemma twice.

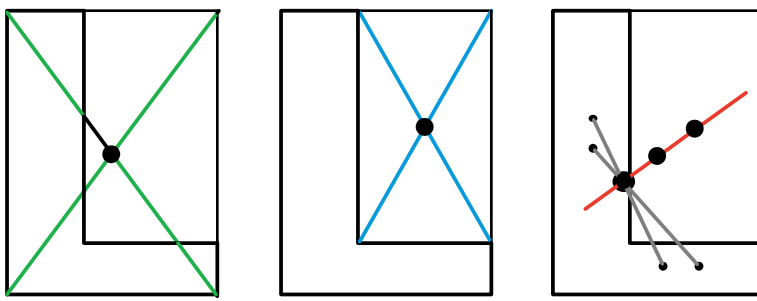


Figure 12.3: Alternative method for finding the centroid of an L-shaped lamina.

Figure 12.4a shows a body APB consisting of two legs PA and PB. Its centroid lies on the segment joining the midpoints M and N of the legs.

If the legs have equal length the figure is symmetric and its centroid is halfway between M and N.

If the lengths are unequal, let PA be the shorter leg and divide body APB into two parts, a symmetric part APA', where PA' has length PA, plus a residual segment $A'B$. The centroid of $A'B$ is at its midpoint $\boldsymbol{p}$, and the symmetric part APA' has its centroid at the point $\boldsymbol{s}$ midway between the centers M and M' of the equal sides, as shown in Figure 12.4b. By the Archimedes Lemma, the centroids $\boldsymbol{p}$ and $\boldsymbol{s}$ lie on the line that contains the centroid $\boldsymbol{c}$ of the compound body APB. This line intersects segment MN at $\boldsymbol{c}$.

Example 4 (An alternative method for finding the centroid in Example 3). An alternative method for determining the centroid $\boldsymbol{c}$ leads to a surprising and useful connection with the triangle MON whose vertices are the midpoints of the sides of triangle APB, shown in Figure 12.5. Instead of dividing APB into a symmetric part plus an extra segment as was done in Example 3, we extend the shorter leg AP to point Q so that AQ has the same length as PB. The composite body consisting of segments AQ and PB of equal length has its centroid at point

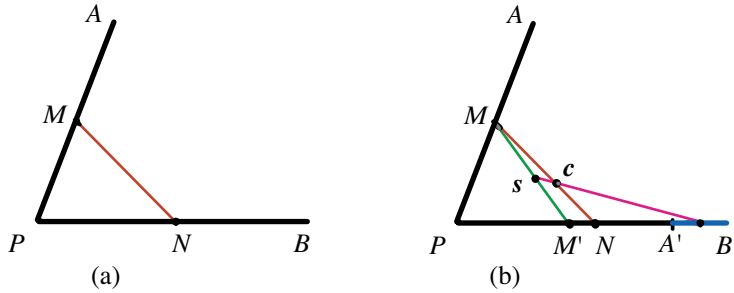


Figure 12.4: Determining the centroid of two intersecting line segments.

$\boldsymbol{t}$, halfway between the midpoints T of AQ and N of PB. Now let $\boldsymbol{e}$ denote the midpoint of PQ. By the variation of the Archimedes Lemma, the centroid $\boldsymbol{c}$ of APB lies on the line joining the centroids $\boldsymbol{e}$ and $\boldsymbol{t}$. This line intersects the segment MN at $\boldsymbol{c}$.

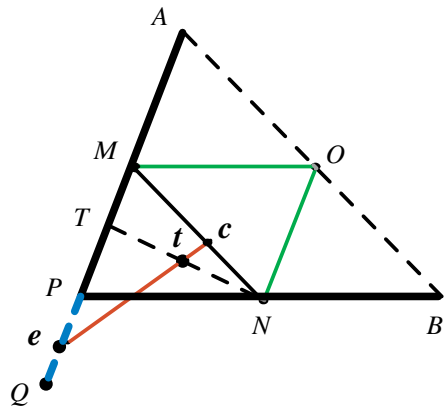


Figure 12.5: Alternative method for finding the centroid of APB.

Example 5 (Yet another method for finding the centroid in Example 4). Now we show that the line through $\boldsymbol{e}$, $\boldsymbol{t}$, and $\boldsymbol{c}$ in Figure 12.5 also passes through the midpoint O of AB. Quadrilateral $PNOM$ is a parallelogram because MO is parallel to PB and ON is parallel to AP. First we show that line $\boldsymbol{e}O$ bisects angle MON, then we show that line $\boldsymbol{e}O$ also contains $\boldsymbol{t}$ and $\boldsymbol{c}$.

Draw a line through $\boldsymbol{e}$ parallel to PN and extend ON until it intersects this line at R, say, as shown in Figure 12.6. We get a new parallelogram $\boldsymbol{e}ROM$ that is also a rhombus (all four of its sides are equal). The reason for this is that

$$\boldsymbol{e}R = PN = \frac{1}{2}PB = \frac{1}{2}AQ,$$

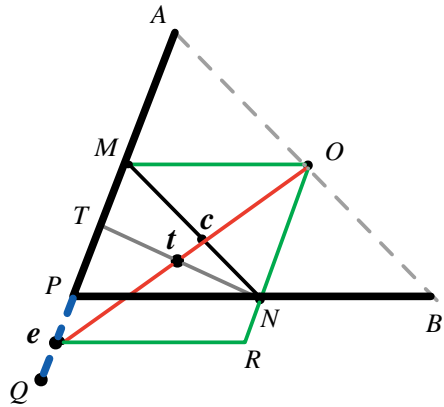


Figure 12.6: The centroid $\boldsymbol{c}$ lies on MN and also on the bisector of angle MON.

and

$$M\boldsymbol{e} = MP + P\boldsymbol{e} = \frac{1}{2}AP + \frac{1}{2}PQ = \frac{1}{2}AQ = \boldsymbol{e}R.$$

Hence $\boldsymbol{e}O$ is a diagonal of the rhombus so it bisects angle MON. But segment TN divides the rhombus into two congruent quadrilaterals $MONT$ and $\boldsymbol{e}RNT$, so $\boldsymbol{t}$, the midpoint of TN, is at the center of the rhombus. Hence diagonal $\boldsymbol{e}O$ passes through $\boldsymbol{t}$ and hence also through $\boldsymbol{c}$ because $\boldsymbol{e}$, $\boldsymbol{t}$, and $\boldsymbol{c}$ lie on a line. Therefore, we have another way to find the centroid $\boldsymbol{c}$ that does not require extending the leg AP. Complete the triangle having PA and PB as two edges, and draw the triangle MON joining the midpoints of its sides. The intersection of MN and the bisector of angle MON is the centroid $\boldsymbol{c}$.

Example 6 (The centroid of the boundary of a triangle). The boundary of a triangle can be regarded as an object made from a uniform wire. Its centroid can be determined by applying the Archimedes Lemma twice. First imagine the boundary of the triangle as a composite body consisting of two line segments meeting at one of the vertices plus a second body consisting of the edge opposite this vertex. This can be done in three ways, as illustrated in Figures 12.7a, b, and c. In each case, the centroid of the first body can be found as in Example 5, and the centroid of the second body lies on the angle bisector joining these two, shown as a red line in each of Figures 12.7a, b, c. The three angle bisectors will intersect at the centroid of the triangular boundary.

Example 7 (The centroid of a triangular lamina). Archimedes was the first to determine the centroid of a triangular lamina T (T includes the boundary of the triangle and all points inside). He found that the centroid of T is at the point of intersection of the three medians.

In fact, Archimedes gave two proofs, both using argument by contradiction. However, both proofs are complicated, involving a large number of auxiliary construction lines, and are not easy to follow. We give a simple direct proof.

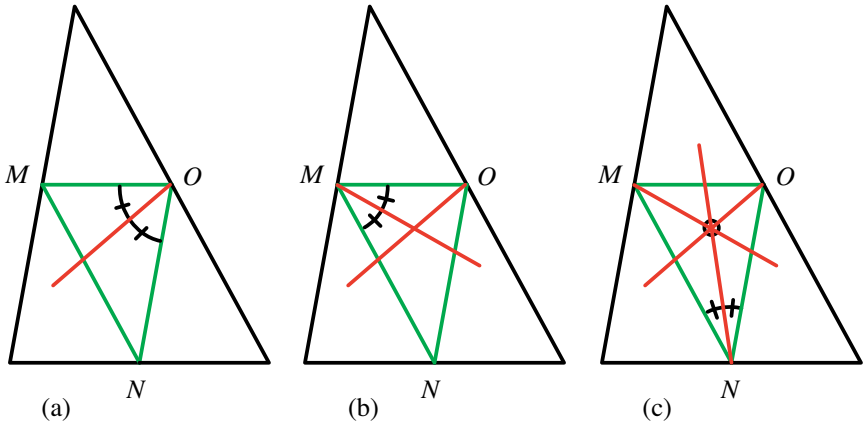


Figure 12.7: The centroid of the boundary of a triangle is at the intersection of the three angle bisectors of the medial triangle.

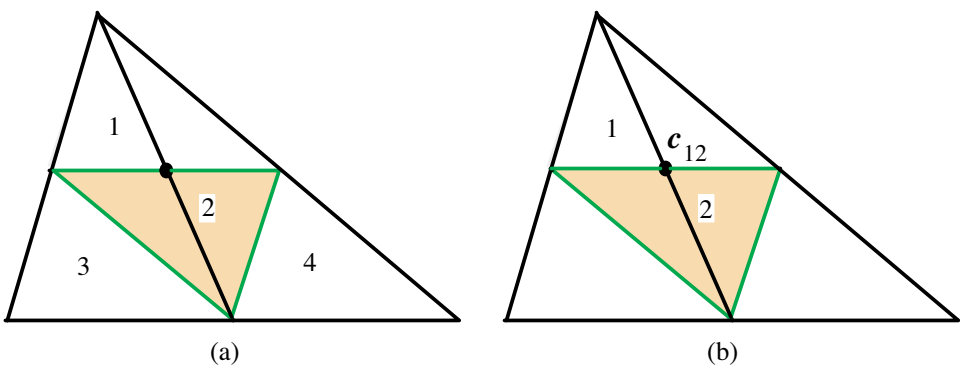


Figure 12.8: A triangle divided by its midlines into four congruent triangles.

Dissect T into four similar triangles by its midlines, as shown in Figure 12.8a. The four pieces are congruent and each edge is half that of the corresponding edge of T. The pieces numbered 3 and 4 are translates of triangle 1, whereas piece number 2 in the center is obtained by rotating triangle 1 through 180^o about the center of their common edge. Triangles 1 and 2 together form a parallelogram whose centroid $\boldsymbol{c}_{12}$ is at its center, which is at the midpoint of the common base of triangles 1 and 2 and also on the median of both triangles 1 and T, as shown in Figure 12.8b.

Now refer to Figure 12.9. Let $\boldsymbol{r}$ be a vector from the uppermost vertex A of T to a point alleged to be the centroid $\boldsymbol{c}_1$ of triangle 1, and let $\boldsymbol{m}$ be the vector (along the median) from A to the midpoint of the opposite side of triangle 1. We wish to prove that $\boldsymbol{r}$ and $\boldsymbol{m}$ are parallel. By the translation property of centroids, the terminal points of the translated vectors in Figure 12.9a represent the centroids $\boldsymbol{c}_3$ and $\boldsymbol{c}_4$ of triangles 3 and 4, respectively. Since triangle 4 is a translation of

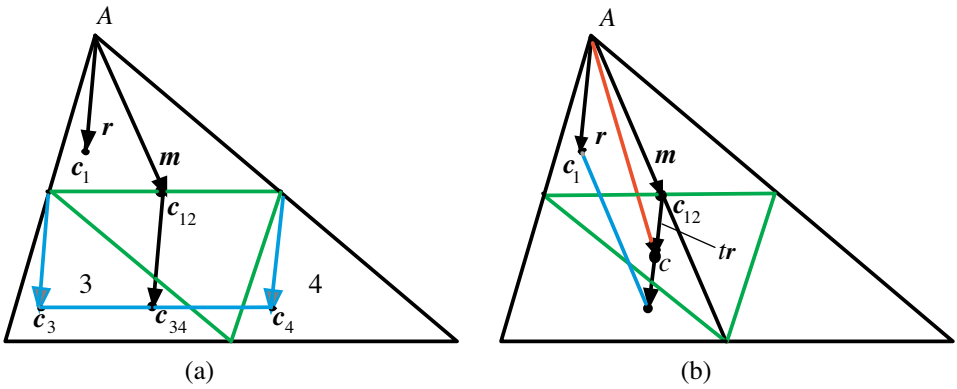


Figure 12.9: Determining the centroid of a triangular lamina.

triangle 3, property 5 tells us that the centroid $\boldsymbol{c}_{34}$ of triangles 3 and 4 together lies midway between $\boldsymbol{c}_3$ and $\boldsymbol{c}_4$. Now the large triangle T is divided into two parts, the parallelogram with centroid $\boldsymbol{c}_{12}$ and the two triangles 3 and 4 with centroid $\boldsymbol{c}_{34}$, so by the Archimedes Lemma the centroid $\boldsymbol{c}$ of T is on the line segment joining $\boldsymbol{c}_{12}$ and $\boldsymbol{c}_{34}$ which, in turn, is parallel to and equal in length to the vector $\boldsymbol{r}$. Hence the vector from $\boldsymbol{c}_{12}$ to $\boldsymbol{c}$ is $t\boldsymbol{r}$, where t lies between 0 and 1, as shown in Figure 12.9b. Therefore the vector from A to $\boldsymbol{c}$ is the vector sum $\boldsymbol{m} + t\boldsymbol{r}$. But triangle 1 can be obtained from T by scaling distances from vertex A by the factor $\frac{1}{2}$, so by the scaling property of centroids we have $\boldsymbol{r} = \frac{1}{2}(\boldsymbol{m} + t\boldsymbol{r})$, which implies $\boldsymbol{m} = (2-t)\boldsymbol{r}$. This shows that $\boldsymbol{r}$ and $\boldsymbol{m}$ are parallel because $\boldsymbol{m}$ is a positive scalar times $\boldsymbol{r}$.

Hence the two centroids $\boldsymbol{c}_1$ and $\boldsymbol{c}$ are on the median from vertex A to the opposite side. Applying the same argument to the other vertices we see that $\boldsymbol{c}$ is on each of the three medians. This shows that the medians of a triangle intersect at the centroid of the triangle. As already shown in Example 1, this point is also the centroid of the three vertices, and the distance from each vertex to the centroid is two-thirds the length of the median.

The authors have devised a variation of this proof that the reader may wish to discover independently. It involves dividing the large triangle into nine congruent triangles by trisecting of its edges. The corner triangles are translates of each other. The lamina is made up of the three corner triangles, together with a centrally symmetric hexagon such that each pair of opposite edges are of equal length. One fringe benefit of this modified proof is that it not only locates the centroid but at the same time gives the distance of the centroid from each vertex.

Example 8 (The centroid of a circular arc and sector). The Archimedes Lemma can also be used to find centroids of curved figures. For example, a standard calculus exercise is to show that the centroid of a circular arc of radius r that subtends a central angle of 2α radians lies along the radius that bisects the arc at a distance $r(\sin\,\alpha)/\alpha$ from the center. When $\alpha = 0$ this distance is to be interpreted

as r, which agrees with the fact that

$$\frac{\sin\alpha}{\alpha} \to 1 \text{ as } \alpha \to 0.$$

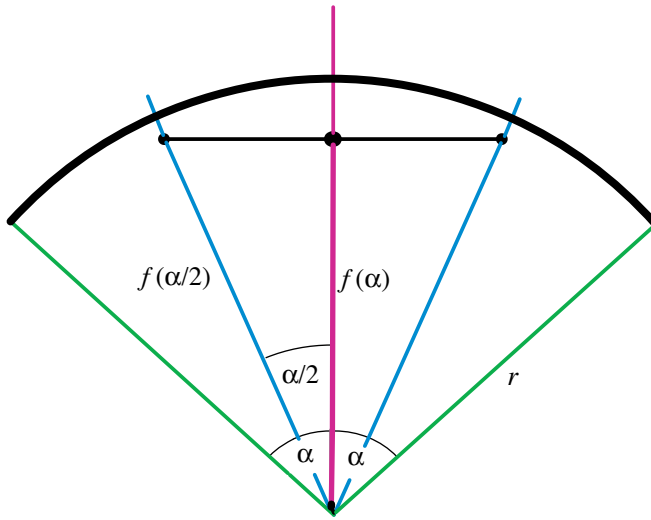


Figure 12.10: Here $f(\alpha)$ denotes the distance from the center of the circle to the centroid of the arc subtended by an angle of 2α radians.

This result can be deduced from the basic properties of centroids. Refer to Figure 12.10, which shows a circular arc of radius r with central angle 2α radians. By symmetry, the centroid lies on the radial line that bisects this arc. Let $f(\alpha)$ denote the distance from the center of the circle to the centroid. The bisector divides the arc into two symmetric arcs of α radians, and the centroids of the smaller arcs are located at a distance $f(\alpha/2)$ from the center. By the Archimedes Lemma, the centroid of the larger arc 2α must lie on the line segment joining the centroids of the smaller arcs. This line segment is perpendicular to the radial line from the center to the centroid of the larger arc. Therefore, from the right triangle in Figure 12.10 we see that

$$f(\alpha) = f(\alpha/2)\cos(\alpha/2)$$

or, replacing α by 2α, we find that

$$f(2\alpha) = f(\alpha)\cos\alpha, \tag{12.4}$$

with $f(0) = r$.

This is a functional equation relating $f(2\alpha)$ with $f(\alpha)$. We will show that the only function continuous at 0 that satisfies (12.4) is $f(\alpha) = r(\sin\alpha)/\alpha$, as required.

Solution of the functional equation (12.4). The function f_0, defined by

$$f_0(\alpha) = r\frac{\sin\alpha}{\alpha} \quad \text{if } \alpha \neq 0,$$

$$f_0(0) = r,$$

satisfies (12.4) because $\sin(2\alpha) = 2\sin\alpha\cos\alpha$. Also, f_0 is continuous at 0. Now let f be a solution of (12.4) that is continuous at 0, and define

$$g(\alpha) = f(\alpha)/f_0(\alpha).$$

We wish to prove that g is the constant function 1. Applying the functional equation (12.4) with $\alpha \neq 0$ we find

$$g(2\alpha) = \frac{f(2\alpha)}{f_0(2\alpha)} = \frac{f(\alpha)\cos(\alpha)}{r\sin(2\alpha)/(2\alpha)}$$

$$= \frac{2\alpha f(\alpha)\cos(\alpha\ )}{2r\sin(\alpha)\cos(\alpha)} = \frac{\alpha f(\alpha)}{r\sin(\alpha)} = g(\alpha).$$

Using this repeatedly we obtain $g(2^n\alpha) = g(\alpha)$ for $n = 1, 2, \ldots$. If $x = 2^n\alpha$ this last relation states that $g(x) = g(x/2^n)$ for $n = 1, 2, \ldots$. Letting $n \to \infty$ and invoking continuity at 0 we find $g(x) = g(0) = 1$ for all x, hence $f(\alpha) = f_0(\alpha)$, as asserted.

A similar argument shows that the centroid of a circular sector of radius r subtending a central angle 2α lies on the radial bisector sector at a distance equal to $(2/3)r(\sin\alpha)/\alpha$ from the center. The Archimedes Lemma leads to the same functional equation, but in this case the limiting value as $\alpha \to 0$ is $f(0) = (2/3)r$. In fact, the method of Example 8 works in some cases of nonuniform mass distribution with radial symmetry, the centroidal distance being $cr(\sin\alpha)/\alpha$, where c is a constant depending on the mass distribution. Example 8 will be solved by another method in Section 13.9, Example 2.

PART 2: CENTROID OF n POINTS

12.5 CENTROIDS CONSTRUCTED GRAPHICALLY

Next we describe two methods for locating the centroid of a finite number of points in 1-space, 2-space, or 3-space by graphical construction, without using coordinates or numerical calculations. The first involves making a guess and forming a closed polygon. The second is an inductive procedure that combines centroids of two disjoint sets to determine the centroid of their union. For 1-space and 2-space both methods can be applied in practice with drawing instruments or with computer graphics programs. Centroids of points in higher-dimensional spaces can be determined with the help of geometric methods by projecting the points onto lower-dimensional spaces, depending on how the points are given. For example, points in 4-space with coordinates (x, y, z, t) can be projected onto points in the xy plane and in the zt plane where graphic methods apply.

Our geometric methods are best illustrated when the weights are equal (determining centroids), and we will indicate in appropriate places how the methods can

be modified to the more general case of unequal weights (determining centers of mass).

Let $\boldsymbol{c}_k = \boldsymbol{r}_k - \boldsymbol{c}$, the geometric vector from centroid $\boldsymbol{c}$ to $\boldsymbol{r}_k$. Then $\boldsymbol{c}_1,\ldots,\boldsymbol{c}_n$ have centroid $\boldsymbol{O}$ because (12.3) shows that

$$\sum_{k=1}^{n} \boldsymbol{c}_k = \boldsymbol{O}. \tag{12.5}$$

Figure 12.11a shows four points and the geometric vectors represented by arrows emanating from their centroid. Figure 12.11b shows the vectors being added head to tail to form a closed polygon. Of course, the vectors can be added in any order.

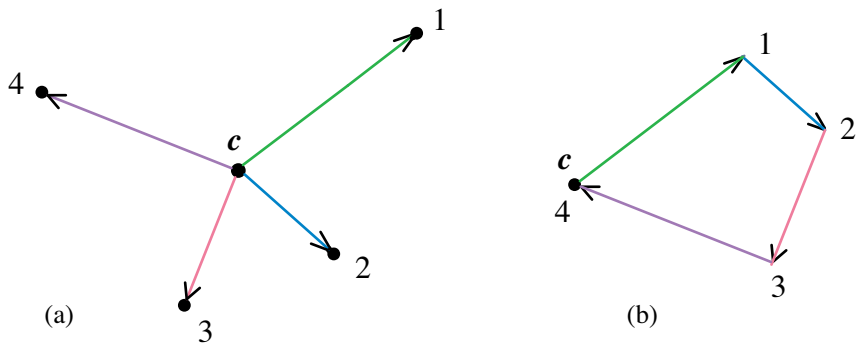


Figure 12.11: Vectors from the centroid form a closed polygon when placed successively head to tail.

Method 1: Closing a polygon. This geometric method for locating the centroid of n points $\boldsymbol{r}_1,\ldots,\boldsymbol{r}_n$ involves choosing a point $\boldsymbol{g}$ as a guess for the centroid. Then we construct a closed polygon and modify the guess once to determine $\boldsymbol{c}$.

We introduce the deficiency vector defined by

$$\boldsymbol{d} = \sum_{k=1}^{n} (\boldsymbol{r}_k - \boldsymbol{g}) = n\boldsymbol{c} - n\boldsymbol{g},$$

as a measure of the deviation between $\boldsymbol{g}$ and $\boldsymbol{c}$. A knowledge of $\boldsymbol{d}$ gives $\boldsymbol{c}$, because

$$\boldsymbol{c} = \boldsymbol{g} + \frac{1}{n}\boldsymbol{d}. \tag{12.6}$$

The vector $\boldsymbol{d}/n$ is the error vector. It tells us exactly what should be added to the guess $\boldsymbol{g}$ to obtain the centroid $\boldsymbol{c}$. Figure 12.12 illustrates the method by an example. The four points used in Figure 12.11a are also shown in Figure 12.12a with a guess $\boldsymbol{g}$ for their centroid, and geometric vectors $\boldsymbol{r}_k - \boldsymbol{g}$ drawn from $\boldsymbol{g}$ to the four points. For simplicity, in the figure we use labels $k = 1, 2, 3, 4$ to denote these vectors. In Figure 12.12b the vectors are placed successively head to tail, starting from $\boldsymbol{g}$

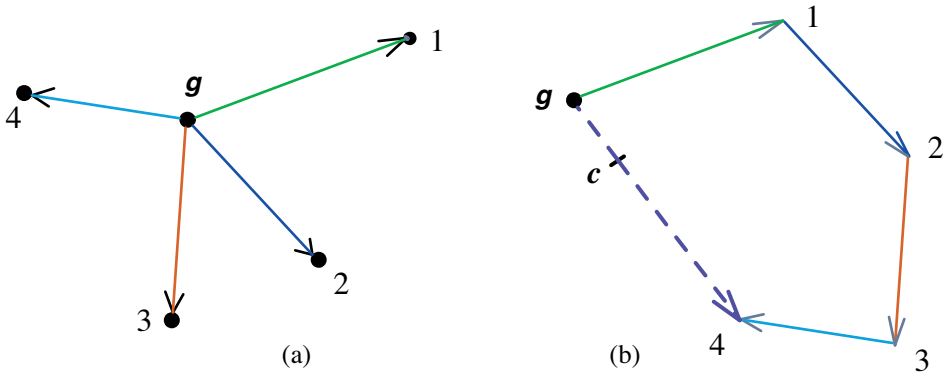


Figure 12.12: If $\boldsymbol{g}$ is not the centroid, the polygon obtained by adding the vectors is not closed. It can be closed by adding $-\boldsymbol{d}$ to the other vectors.

to form the sum $\boldsymbol{d}$. If a lucky guess placed $\boldsymbol{g}$ at the centroid, the vectors placed head to tail would form a closed polygon as in Figure 12.11b, and $\boldsymbol{d}$ would be zero. But in Figure 12.12a, $\boldsymbol{g}$ is not the centroid, and the polygon formed by the four vectors in Figure 12.12b is not closed. However, an additional vector joining the tail of $\boldsymbol{r}_1 - \boldsymbol{g}$ to the head of $\boldsymbol{r}_4 - \boldsymbol{g}$ will close the polygon. This vector, shown as a broken line in Figure 12.12b, is the deficiency vector $\boldsymbol{d}$. We find $\boldsymbol{c}$ by adding the error vector $\boldsymbol{d}/4$ to $\boldsymbol{g}$. In practice, Figures 12.12a and 12.12b can be drawn on the same graph. We have separated them here for the sake of clarity.

Although this example illustrates the method for four points in a plane, the method works equally well for any number of points in 1-space, 2-space, or 3-space. For n points we add the error vector $\boldsymbol{d}/n$ to $\boldsymbol{g}$, as indicated by (12.6), to get the centroid $\boldsymbol{c}$. The error vector $\boldsymbol{d}/n$ is easily constructed geometrically. In fact, to multiply $\boldsymbol{d}$ by a positive scalar λ, plot $1/\lambda$ on a number line drawn in a convenient direction not parallel to $\boldsymbol{d}$, and join $1/\lambda$ and the head of $\boldsymbol{d}$ with a line segment. A parallel line to $\boldsymbol{d}$ through the unit on the number line intersects $\boldsymbol{d}$ at $\lambda\boldsymbol{d}$, as is easily verified by similarity of triangles.

Multiplication by a general positive scalar is needed when the method of guessing once is used to find the weighted average $\boldsymbol{c}$ of n given points as defined by (12.1). As before, we make a guess $\boldsymbol{g}$ and let

$$\boldsymbol{d} = \sum_{k=1}^{n} w_k(\boldsymbol{r}_k - \boldsymbol{g}) = W_n(\boldsymbol{c} - \boldsymbol{g}).$$

This gives us $\boldsymbol{c} = \boldsymbol{g} + \boldsymbol{d}/W_n$, and it is easily verified that the foregoing geometric method can be adapted to find $\boldsymbol{d}$, the error vector $\boldsymbol{d}/W_n$, and hence $\boldsymbol{c}$. Two simple examples illustrate how the method yields familiar interpretations of the centroid.

Example 9 (Centroid of two points). The centroid of two points is midway between them, and it is instructive to see how the geometric method works in this simple case.

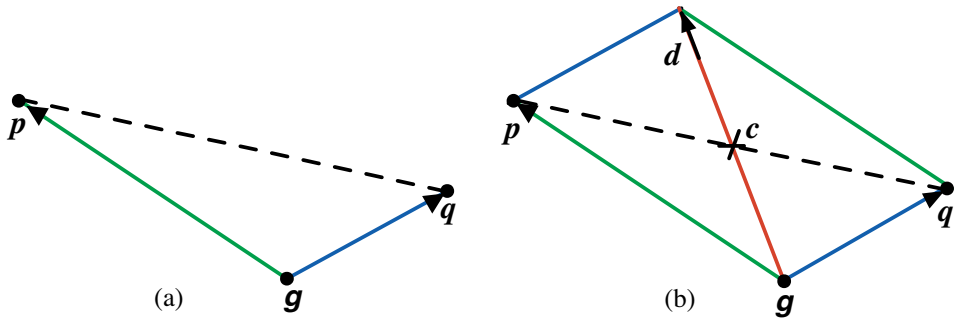


Figure 12.13: Determining the centroid of two points.

Figure 12.13a shows two points $\boldsymbol{p}$ and $\boldsymbol{q}$ and a guess $\boldsymbol{g}$ for their centroid. This may seem to be an outlandish guess because $\boldsymbol{g}$ is not on the line through $\boldsymbol{p}$ and $\boldsymbol{q}$. Nevertheless, the method works no matter where $\boldsymbol{g}$ is chosen. When we add the vector from $\boldsymbol{g}$ to $\boldsymbol{p}$ to that from $\boldsymbol{g}$ to $\boldsymbol{q}$, the sum

$$\boldsymbol{d} = (\boldsymbol{p} - \boldsymbol{g}) + (\boldsymbol{q} - \boldsymbol{g})$$

is one diagonal of a parallelogram, as shown in Figure 12.13b, and Eq. (12.6) tells us that the centroid is given by

$$\boldsymbol{c} = \boldsymbol{g} + \boldsymbol{d}/2 = (\boldsymbol{p} + \boldsymbol{q})/2,$$

which lies midway between $\boldsymbol{p}$ and $\boldsymbol{q}$.

Actually, by choosing $\boldsymbol{g}$ outside the line through $\boldsymbol{p}$ and $\boldsymbol{q}$, there is no need to construct $\boldsymbol{d}/2$, because we know that $\boldsymbol{c}$ lies on this line at the point where $\boldsymbol{d}$ intersects it.

Example 10 (Centroid of three points). Choose three points $\boldsymbol{p}$, $\boldsymbol{q}$, $\boldsymbol{r}$, and make a guess $\boldsymbol{g}$ for their centroid (Figure 12.14a). The points are vertices of a triangle, possibly degenerate, but the guess $\boldsymbol{g}$ need not be in its the plane. According to Method 1, we form the sum

$$\boldsymbol{d} = (\boldsymbol{p} - \boldsymbol{g}) + (\boldsymbol{q} - \boldsymbol{g}) + (\boldsymbol{r} - \boldsymbol{g}) = (\boldsymbol{p} + \boldsymbol{q} + \boldsymbol{r}) - 3\boldsymbol{g}$$

and by (12.6) the centroid is

$$\boldsymbol{c} = \boldsymbol{g} + \boldsymbol{d}/3 = (\boldsymbol{p} + \boldsymbol{q} + \boldsymbol{r})/3.$$

It should be noted that the method works even if $\boldsymbol{p}$, $\boldsymbol{q}$, $\boldsymbol{r}$ are collinear provided we choose $\boldsymbol{g}$ not on the same line, which turns out to be a wise guess in this case. For such a $\boldsymbol{g}$, we know that $\boldsymbol{c}$ is on the line through $\boldsymbol{p}$, $\boldsymbol{q}$, $\boldsymbol{r}$, so the vector $\boldsymbol{d}$ will automatically intersect the line at $\boldsymbol{c}$, and hence there is no need to divide $\boldsymbol{d}$ by 3.

We can also verify that in the nondegenerate case all three medians of the triangle meet at the centroid, and that the centroid divides each median in the ratio 2:1.

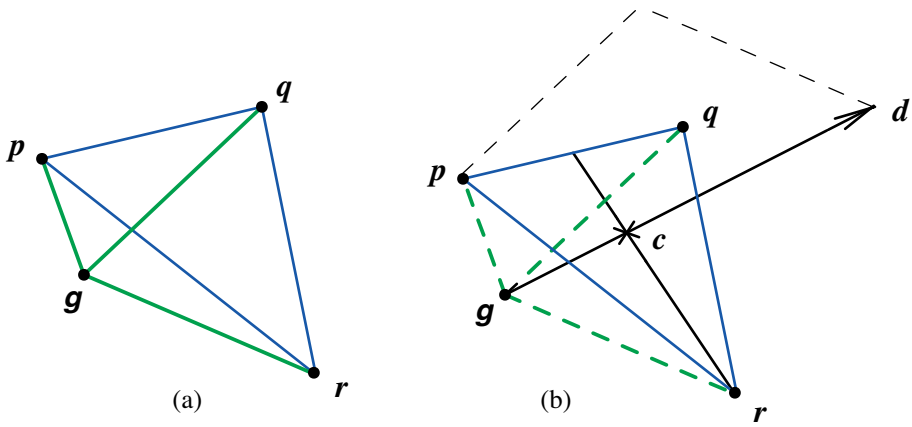


Figure 12.14: The three medians of a triangle intersect at the centroid of the vertices.

In Figure 12.14b the centroid is placed at the origin so that $\boldsymbol{p}+\boldsymbol{q}+\boldsymbol{r}=\boldsymbol{O}$. First consider the median from $\boldsymbol{r}$ to the edge joining vertices $\boldsymbol{p}$ and $\boldsymbol{q}$. The vector from the centroid to the midpoint of this edge is $(\boldsymbol{p}+\boldsymbol{q})/2$ whereas $\boldsymbol{r}=-(\boldsymbol{p}+\boldsymbol{q})$. This shows that the median passes through the centroid and that the distance from the centroid to vertex $\boldsymbol{r}$ is twice that from the centroid to the midpoint of the opposite edge. By interchanging symbols, we find the same is true for the other two medians. So all three medians meet at the centroid, and the centroid divides each median in the ratio 2:1 as asserted.

Variation of method 1 that avoids dividing.

Figure 12.15 shows an alternative way to determine the centroid $\boldsymbol{c}$ of the points in Figure 12.12 that avoids constructing the error vector $\boldsymbol{d}/4$. Figure 12.15a shows the

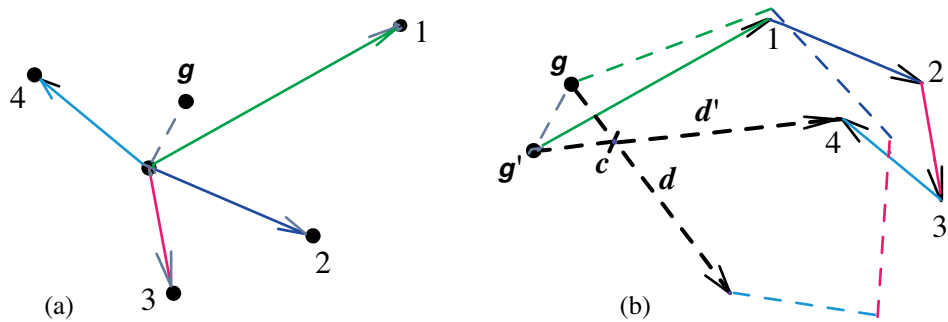


Figure 12.15: The centroid as the intersection of two deficiency vectors $\boldsymbol{d}$ and $\boldsymbol{d}'$.

same four points with a new guess $\boldsymbol{g}'$ chosen not on the line through $\boldsymbol{d}$. The polygon of Figure 12.12b appears again in Figure 12.15b as dashed lines, together with a

new polygon (solid lines) obtained by adding the vectors $\boldsymbol{r}_k - \boldsymbol{g}$. This produces a new deficiency vector $\boldsymbol{d}'$ joining the tail of $\boldsymbol{r}_1 - \boldsymbol{g}'$ to the head of $\boldsymbol{r}_4 - \boldsymbol{g}'$. Because $\boldsymbol{g}'$ is not on the line through $\boldsymbol{d}$, the two geometric vectors $\boldsymbol{d}$ and $\boldsymbol{d}'$ intersect at $\boldsymbol{c}$, as shown by the example in Figure 12.15b.

Although the example in Figure 12.15 treats four points, the method also works for any finite number of points. It should be noted that if the points are collinear a second guess is not needed provided we choose the first guess $\boldsymbol{g}$ not on the line through them. The construction used for three collinear points in Example 10 also works for any number of collinear points. The deficiency vector $\boldsymbol{d}$ will intersect the line through these points at the centroid $\boldsymbol{c}$.

The case of unequal weights is treated similarly as described previously. Make two guesses, and the two deficiency vectors so constructed can be shown to intersect at $\boldsymbol{c}$.

Method 2: Inductive process. This method regards the given set of points as the union of two disjoint subsets whose centroids are known or can be easily determined. The centroids of the subsets are combined to determine the centroid of the union. The process depends on how the subsets are selected. For example, we can find the centroid of $n+1$ points if we know the centroid of any n of them. If $\boldsymbol{c}_n$ denotes the centroid of points $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n$, so that

$$\boldsymbol{c}_n = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{r}_k,$$

then

$$\boldsymbol{c}_{n+1} = \frac{1}{n+1} \sum_{k=1}^{n+1} \boldsymbol{r}_k = \frac{1}{n+1}\Big(\sum_{k=1}^{n} \boldsymbol{r}_k + \boldsymbol{r}_{n+1}\Big),$$

or

$$\boldsymbol{c}_{n+1} = \frac{1}{n+1}(n\boldsymbol{c}_n + \boldsymbol{r}_{n+1}). \tag{12.7}$$

In other words, $\boldsymbol{c}_{n+1}$ is a weighted average of the two points $\boldsymbol{c}_n$ and $\boldsymbol{r}_{n+1}$, with weight n attached to $\boldsymbol{c}_n$ and weight 1 attached to $\boldsymbol{r}_{n+1}$. Because $\boldsymbol{c}_{n+1}$ is a convex combination of $\boldsymbol{c}_n$ and $\boldsymbol{r}_{n+1}$, it lies on the line joining $\boldsymbol{c}_n$ and $\boldsymbol{r}_{n+1}$. Moreover, from (12.7) we find

$$\boldsymbol{c}_{n+1} - \boldsymbol{c}_n = \frac{1}{n+1}(\boldsymbol{r}_{n+1} - \boldsymbol{c}_n),$$

which shows that the distance between $\boldsymbol{c}_{n+1}$ and $\boldsymbol{c}_n$ is $1/(n+1)$ times the distance between $\boldsymbol{r}_{n+1}$ and $\boldsymbol{c}_n$. Repeated use of (12.7) provides a method for determining the centroid of any finite set.

In the case of unequal weights, (12.7) becomes

$$\boldsymbol{c}_{n+1} = \frac{1}{W_{n+1}}(W_n\boldsymbol{c}_n + w_{n+1}\boldsymbol{r}_{n+1}),$$

a convex combination of $\boldsymbol{c}_n$ and $\boldsymbol{r}_{n+1}$ that lies on the line joining $\boldsymbol{c}_n$ and $\boldsymbol{r}_{n+1}$. This implies

$$\boldsymbol{c}_{n+1} - \boldsymbol{c}_n = \frac{w_{n+1}}{W_{n+1}}(\boldsymbol{r}_{n+1} - \boldsymbol{c}_n),$$

so the distance from $\boldsymbol{c}_{n+1}$ to $\boldsymbol{c}_n$ is w_{n+1}/W_{n+1} times that between $\boldsymbol{r}_{n+1}$ and $\boldsymbol{c}_n$.

Example 11 (Centroid of five points). Figure 12.16 shows how this method yields the centroid of five points. Let $\boldsymbol{c}_k$ denote the centroid of the set of points 1, 2, ..., k. The centroid $\boldsymbol{c}_1$ of point 1 is, of course, the point itself. Using (12.7) with $n = 1$ we find $\boldsymbol{c}_2$ is midway between 1 and 2. Now connect $\boldsymbol{c}_2$ with point 3 and divide the distance between them by 3 to find the centroid $\boldsymbol{c}_3$. Then connect $\boldsymbol{c}_3$ with point 4 and divide the distance between them by 4 to find $\boldsymbol{c}_4$. Finally, connect $\boldsymbol{c}_4$ with point 5 and divide the distance between them by 5 to get $\boldsymbol{c}_5$. It is clear that this method will work for any number of points.

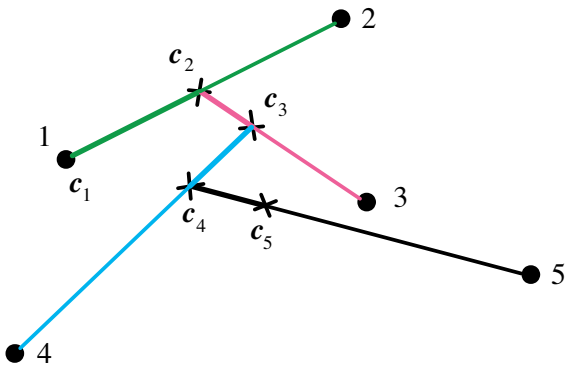


Figure 12.16: Distance between $\boldsymbol{c}_{k+1}$ and $\boldsymbol{c}_k$ is $1/(k+1)$ times the distance between $\boldsymbol{r}_{k+1}$ and $\boldsymbol{c}_k$.

Variation of Method 2 using only bisection and connecting points.

A variation of Method 2 can be used by those who prefer not to divide vectors into more than two equal parts. This construction uses only two geometric operations–bisection of segments, and connecting points–so it applies only to the case of equal weights. We begin with a special example that constructs the centroid using only repeated bisection of segments.

Example 12 (Centroid of four points). The centroid of four points $\boldsymbol{p}$, $\boldsymbol{q}$, $\boldsymbol{r}$, $\boldsymbol{s}$, is given by

$$\boldsymbol{c} = (\boldsymbol{p} + \boldsymbol{q} + \boldsymbol{r} + \boldsymbol{s})/4.$$

By writing this in the form

$$\boldsymbol{c} = \frac{1}{2}\left(\frac{\boldsymbol{p} + \boldsymbol{q}}{2} + \frac{\boldsymbol{r} + \boldsymbol{s}}{2}\right), \tag{12.8}$$

we see that the centroid is at the midpoint of the segment joining $(\boldsymbol{p} + \boldsymbol{q})/2$ and $(\boldsymbol{r} + \boldsymbol{s})/2$ which, in turn, are midpoints of the segments from $\boldsymbol{p}$ to $\boldsymbol{q}$ and from $\boldsymbol{r}$ to $\boldsymbol{s}$. By permuting symbols in (12.8), we find the centroid is also the midpoint of

the segment joining $(\boldsymbol{s}+\boldsymbol{p})/2$ and $(\boldsymbol{q}+\boldsymbol{r})/2$, and of the segment joining points $(\boldsymbol{p}+\boldsymbol{r})/2$ and $(\boldsymbol{q}+\boldsymbol{s})/2$. Two quadrilaterals with vertices $\boldsymbol{p}$, $\boldsymbol{q}$, $\boldsymbol{r}$, $\boldsymbol{s}$ are shown in Figure 12.17, one convex and one not convex. In each case the segment joining $\boldsymbol{p}$ and $\boldsymbol{r}$ (shown dashed) is a diagonal of the quadrilateral with vertices $\boldsymbol{p}$, $\boldsymbol{q}$, $\boldsymbol{r}$, $\boldsymbol{s}$, as is the segment joining $\boldsymbol{q}$ and $\boldsymbol{s}$. The centroid $\boldsymbol{c}$ lies midway between midpoints of the edges and of the diagonals of the quadrilateral. Any four of the six bisections shown are enough to determine the centroid.

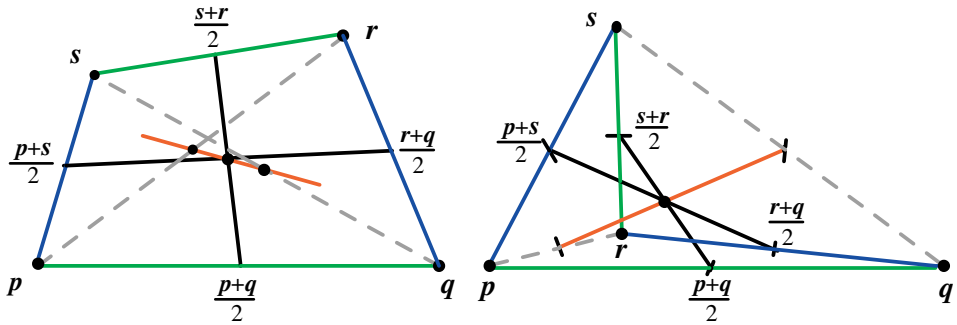


Figure 12.17: For each set of four points the centroid is midway between midpoints of edges and diagonals.

If the four points are collinear, (12.8) shows that three bisections alone suffice to determine their centroid. An obvious iteration of (12.8) shows that, if $k \geq 1$, successive bisection suffices to find the centroid of any $2k$ points, collinear or not. If the number of points is not a power of 2 we will show that again two operations, bisection of segments and connecting points, suffice to find their centroid. One of the principal tools used in this method is the Archimedes Lemma of Section 12.2. When adapted to finite sets of points, this property can be modified as follows:

Archimedes Lemma for Finite Sets. If a finite set with centroid $\boldsymbol{c}$ is divided into two disjoint sets with centroids $\boldsymbol{c_1}$ and $\boldsymbol{c}_2$, then the three centroids are collinear. Moreover, $\boldsymbol{c}$ lies between $\boldsymbol{c_1}$ and $\boldsymbol{c}_2$.

This is easily proved by the same method we used to obtain (12.7) in Method 2. Instead of (12.7) we get a formula of the form

$$\boldsymbol{c} = \frac{1}{n_1 + n_2}(n_1\boldsymbol{c}_1 + n_2\boldsymbol{c}_2),$$

where $\boldsymbol{c}_1$ is the centroid of n_1 points and $\boldsymbol{c}_2$ is the centroid of n_2 points. This shows that $\boldsymbol{c}$ is a convex combination of $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ and hence lies on the line segment joining them.

The next example shows how the Archimedes Lemma can be used to determine centroids of finite sets of points using only bisection of segments and connecting pairs of points.

Example 13 (Centroid of five points). Figures 12.18a and 12.18b show five points distributed in two ways as the union of four points and one point. In each case the centroid of the four points is found as in Example 12, so by the Archimedes Lemma the centroid of all five points lies on the dashed segment joining the centroid with the fifth point. Figure 12.18c shows the intersection of the two dashed segments giving the required centroid of the five points.

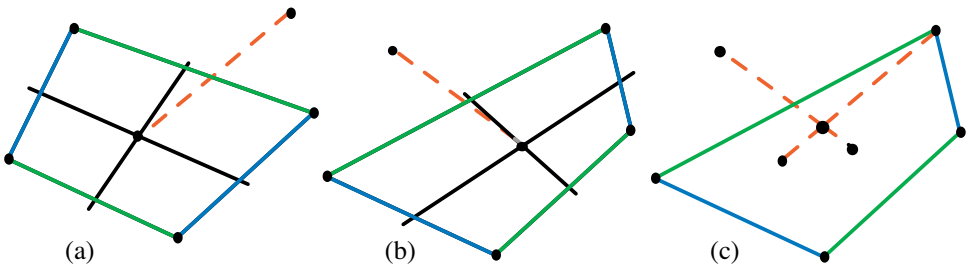


Figure 12.18: Centroid of five points found by using the Archimedes Lemma twice.

In the special case when the five points are collinear, the method cannot give the actual centroid $\boldsymbol{c}$ because the intersecting lines are also collinear. But in this case we can adjoin a sixth point not on the common line and find the centroid $\boldsymbol{c}'$ of the six points by using the Archimedes Lemma twice. Three of the five collinear points together with the sixth point form a quadrilateral whose centroid $\boldsymbol{c}_1$ can be found as in Example 12. The remaining two of the five collinear points have their centroid $\boldsymbol{c}_2$ at their midpoint. By the Archimedes Lemma the centroid $\boldsymbol{c}'$ lies on the line joining $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$. Now repeat the argument, using a different choice of three of the five collinear points to find another line containing $\boldsymbol{c}$. Then the line from the sixth point through $\boldsymbol{c}'$ intersects the line through the five points at their centroid $\boldsymbol{c}$.

Now we have all the ingredients needed to show that centroids can be determined geometrically using only bisection of segments and connecting pairs of points. We state the result as a theorem, whose proof is constructive and outlines a variation of Method 2.

Theorem 12.1. *For $n \geq 2$, the centroid of n points in m-space can be constructed using only bisection of segments and connecting distinct points.*

Proof. The proof is by induction on n. For $n = 2$ bisection suffices. For $n = 3$ we use bisection and drawing lines in the same manner as described in Example 13. For $n = 4$, bisection alone suffices as described in Example 12.

Now suppose the theorem is true for n points, and consider a set of $n+1$ points. Select one of the $n+1$ points and join it to the centroid of the remaining n points which, by the induction hypothesis, has been obtained by bisection of segments and connecting points. Repeat the process, using a different choice for point $n+1$. If all $n+1$ points are not collinear, the two lines so obtained will intersect at their centroid. If all the $n+1$ points are collinear, choose an additional point outside this line and form, in two ways, a set of n points and a disjoint set of two points (as in

Example 13) and apply the inductive procedure twice. This gives two lines whose point of intersection is the centroid, obtained by using only bisection of segments and connecting points.

Now we have several procedures at our disposal for finding centroids by an inductive method, two of which have been illustrated for five points. In Example 11 (Figure 12.16) we advanced one point at a time, and in Example 13 (Figure 12.18) we decomposed the set of five points in two different ways as the disjoint union of a single point and a set of four points. In general we can decompose a set of n points into two disjoint subsets in two different ways and use the Archimedes Lemma twice as was done in Example 13. The choice of subsets is a matter of preference, depending on the number of points.

12.6 ALTERNATIVE BISECTION INDUCTIVE METHOD

An alternative method gives the centroid of any sequence of n points,using exactly $n-1$ bisections of segments, and with each bisection it constructs two lines connecting distinct points. The method can be described inductively as follows. Start with three noncollinear points 1, 2, 3. Point 1 is its own centroid, C_1. The bisector B_1 of the segment joining 1 and 2 is their centroid, C_2. Now bisect the segment joining 2 and 3 to get their bisector, B_2. The segment joining 1 and B_2 intersects that joining 3 and C_2 at the centroid C_3 of points 1, 2, 3. (Figure 12.19a.) Now suppose that centroids $C_3, C_4, ..., C_N$ have been constructed in the same manner for N points, where $N \geq 4$. Choose point $N+1$, noncollinear with N and C_N,

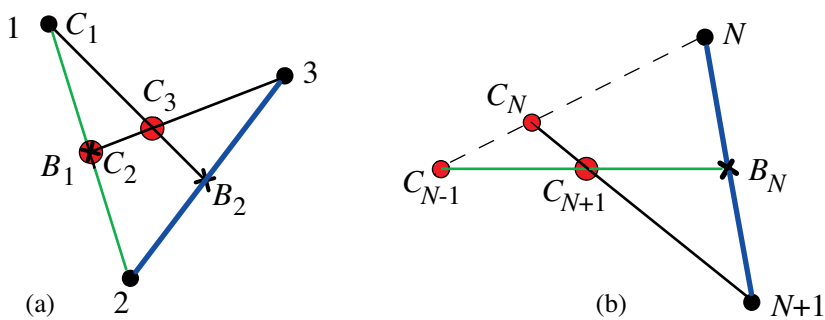


Figure 12.19: Alternative bisection inductive method uses only one new bisection and two additional line segments at each step.

and let B_N denote the bisector of the segment joining N and $N+1$, as shown in Figure 12.19b. The reader can verify that the segment joining B_N with C_{N-1} intersects the segment joining C_N with $N+1$ at the centroid C_{N+1}. This step requires only one new bisection and drawing two additional line segments, and proves our assertion by induction. The advantage of this method is that only three earlier points are needed at each step of the induction. A Java applet of this process is at www.mamikon.com/Centroid.html.

We devised this alternative after receiving a communication from an amateur mathematician, Jason Shields, who sent us elaborate accurate drawings of his constructions for $1024 = 2^{10}$ points and for $1155 = 3 \times 5 \times 7 \times 11$ points, using only bisections and joining of distinct points. His method is based on factorization of the number of points into prime powers.

12.7 GENERALIZATION OF A PUTNAM PROBLEM

We conclude with an extension of Problem A4 of the 29th William Lowell Putnam Mathematical Competition (1968), which asked to show that the sum of the squares of the $n(n-1)/2$ distances between n distinct points on the surface of a unit sphere in 3-space is at most n^2. (See [56].) Several solutions are known, including one by the authors given in Chapter 14 as an application of Theorem 14.3, using a method that reveals the natural role played by the centroid. The same method is used here together with graphic construction of centroids to solve a more interesting and more general problem in m-space.

The generalized problem asks for the maximum value of the sum of squares of all distances among n points $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$, in m-space that lie on concentric spheres of radii $|\boldsymbol{r}_1|, \ldots, |\boldsymbol{r}_n|$. We can imagine an analogous problem in atomic physics where electrons move on concentric spheres and we ask to minimize the potential energy of the system, which requires minimizing the sum of reciprocals of the distances between charges. Here we wish to maximize the sum of the squares of the distances between points. Theorem 14.3 of Chapter 14 shows that the sum is related to their centroid $\boldsymbol{c}$ by the formula

$$\sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k|^2 = n \sum_{k=1}^{n} |\boldsymbol{r}_k|^2 - n^2 |c|^2, \tag{12.9}$$

where $\sum_{k<i}$ is an abbreviation for the double sum $\sum_{i=1}^{n} \sum_{k=1}^{i-1}$. Using (12.9), we can easily maximize the sum on the left, because the right-hand side has its maximum value if and only if $|\boldsymbol{c}|$ reaches its minimum. In other words, locate the points so the centroid is as close as possible to the common center of the spheres. If the value $|\boldsymbol{c}| = 0$ is possible, then $\boldsymbol{c}$ is at the common center (which is chosen also as the origin $\boldsymbol{O}$), and the maximum value is n times the sum of the squares of the radii of the concentric spheres:

$$\max \sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k|^2 = n \sum_{k=1}^{n} |\boldsymbol{r}_k|^2.$$

For example, in the original Putnam problem, each $|\boldsymbol{r}_k| = 1$, and by locating the points so their centroid is at the center, we find that (12.10) gives the required maximum:

$$\max \sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k|^2 = n^2. \tag{12.10}$$

However, if the points are required to lie on different spheres, the problem of locating them to maximize the sum of squares is more difficult because it is not always

possible to place their centroid at the common center. But it can be solved by the graphical method for finding the centroid by closing a polygon. This solution gives a constructive approach as well as a visual interpretation of the results. The solution splits naturally into two cases, depending on how the largest radius $|\boldsymbol{r}_n|$ compares with the sum of all other radii.

Case 1. $|\boldsymbol{r}_n| < \sum_{k=1}^{n-1} |\boldsymbol{r}_k|$, $n \geq 3$. When $n = 3$, this is the triangle inequality, and for $n > 3$ it is a polygonal inequality that makes it possible to choose the vectors $\boldsymbol{r}_1,\ldots, \boldsymbol{r}_n$, to have sum zero. In this case we place the origin at the common center of the spheres, and connect the vectors with hinges, head to tail, to form a closed polygon. Because the vectors joining successive edges have sum zero, they can be translated so all initial points are at the origin. The terminal points $\boldsymbol{r}_1,\ldots,$ $\boldsymbol{r}_n$, will lie on concentric spheres of radii $|\boldsymbol{r}_1|, \ldots, |\boldsymbol{r}_n|$, their centroid will be at the common center, and the points will satisfy the maximal sum relation (12.10).

If $n = 3$ the points are vertices of a rigid triangle. But if $n > 3$, there are infinitely many incongruent solutions represented by closed flexible polygons. Any one of these shapes provides a solution: translate each vector parallel to itself to bring all tails to a common point, the common center of the spheres.

Case 2. $|\boldsymbol{r}_n| \geq \sum_{k=1}^{n-1} |\boldsymbol{r}_k|$, $n \geq 2$. In this case the vectors $\boldsymbol{r}_1,\ldots, \boldsymbol{r}_n$ cannot have sum zero unless $|\boldsymbol{r}_n| = \sum_{k=1}^{n-1} |\boldsymbol{r}_k|$ and the vectors are on a line. In general, we get the largest possible sum of squares of distances from each other by arranging the vectors along a straight line, with the first $n-1$, vectors $\boldsymbol{r}_1,\ldots, \boldsymbol{r}_{n-1}$ pointing in the same direction, and the nth vector $\boldsymbol{r}_n$ (with the largest radius) pointing in the diametrically opposite direction. Unlike in Case 1, this solution is unique; it gives the largest possible sum of squares of distances from each other consistent with (12.9), but it will not reach the maximum provided by the right-hand side of (12.10) because the centroid is not at the origin.

Where is the centroid?

Using the guess $\boldsymbol{g} = \boldsymbol{O}$, we find the deficiency vector $\boldsymbol{d}$ in this case is $\boldsymbol{d} = \sum_{k=1}^{n} \boldsymbol{r}_k = n\boldsymbol{c}$, hence $\boldsymbol{c} = \boldsymbol{d}/n$. Therefore (12.9) implies that the maximum is given by

$$\sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k|^2 = n \sum_{k=1}^{n} |\boldsymbol{r}_k|^2 - |\boldsymbol{d}|^2,$$

because the vectors are along a line.

As in the original Putnam problem, it is surprising that the maximum in both cases is independent of the dimensionality m of the space if $m \geq 2$. Any solution in one common equatorial plane of the spheres (that is, for $m = 2$) is also a solution in all higher-dimensional spaces.

NOTES ON CHAPTER 12

The material in Sections 12.1 through 12.4 was originally published in [6]; that in Sections 12.5 through 12.7 appeared in [14].

Chapter 13

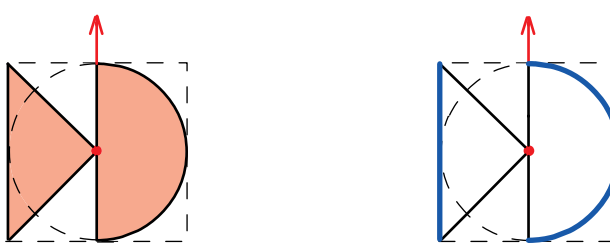
NEW BALANCING PRINCIPLES WITH APPLICATIONS

These problems can be easily solved by the methods developed in this chapter. The reader may wish to try solving them before reading the chapter.

An isosceles right triangle and a semicircular disk are inscribed in a square as depicted below. Show that:

The two regions in (a) *are in area equilibrium about the vertical axis through the center* (*indicated by the arrow*).

The semicircular arc and the base of the triangle in (b) *are in arclength equilibrium about the same axis.*



(a) (b)

Use the foregoing results to show that:

The volume of a sphere is two-thirds that of the right circumscribing cylinder (*whose height is the diameter of the sphere*).

The surface area of a sphere is two-thirds the total surface area of the same circumscribing cylinder.

CONTENTS

13

Archimedes' mechanical balancing methods led him to stunning discoveries concerning the volume of a sphere, and of a cylindrical wedge. This chapter introduces new balancing principles (different from those of Archimedes) including a balance-revolution principle and double equilibrium that go much further. They yield a host of surprising relations involving both volumes and surface areas of circumsolids of revolution, as well as higher-dimensional spheres, cylindroids, spherical wedges, and cylindrical wedges. The concept of cylindroid, introduced here, is crucial for extending to higher dimensions Archimedes' classical relations on the sphere and cylinder. We also provide formulas for centroids of various portions of these objects, including remarkable new results for hemispheres in n-space. Throughout the chapter we adhere to Archimedes' style of reducing properties of complicated objects to those of simpler objects.

13.1 INTRODUCTION

The sphere and circumscribing cylinder engraved on Archimedes' tombstone commemorate his landmark discovery that their volumes and surface areas are related by the same ratio 2/3. He discovered the volume relation and many other geometrical results by mechanical balancing.

In particular, he found the volume of a cylindrical wedge by introducing the balancing shown in Figures 13.1a and 13.1b. First, he uses the Pythagorean theorem to balance the lengths of horizontal chords of a triangle and a semicircular disk with respect to a vertical axis through its center as in Figure 13.1a. He then builds two-dimensional regions from these chords, the right triangle and semicircular disk in Figure 13.1b, which now are in *area balance* with each other. This balancing

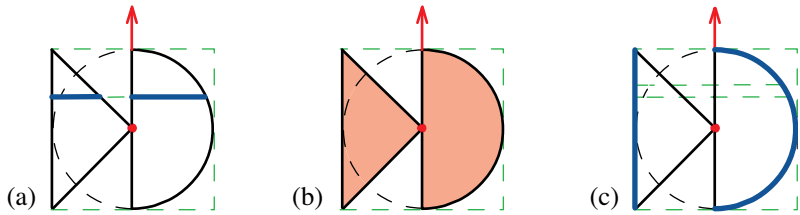


Figure 13.1: Archimedes' balancing of a triangle and semicircular disk: (a) chord-by-chord; (b) by areas. (c) Archimedes would have been pleased to learn that the semicircle and the vertical base of the triangle are in arclength equilibrium as well.

eventually led him to the volume of a cylindrical wedge, as described in [47; *Method*, Prop. 11].

This chapter introduces new balancing principles for treating these problems. The wedge is treated with a new balance-wedge principle introduced in Section 13.4. The sphere and cylinder are treated with a new balance-revolution principle introduced in Section 13.3 and illustrated here by showing that Archimedes' balancing in Figures 13.1a and 13.1b directly yield his volume result in two different ways. (For the volume of a sphere, Archimedes balances completely different objects.)

First, rotation of the chords in Figure 13.1a about the balancing axis, as indicated in Figure 13.2a, produces a circular annulus and a circular disk which, by a theorem of Pappus, have equal areas. Equality of these cross-sectional areas, in turn, yields equality of the volumes of the sphere and punctured cylinder (as was shown in Theorem 5.1), giving Archimedes' 2/3 ratio for the sphere and its circumscribing cylinder.

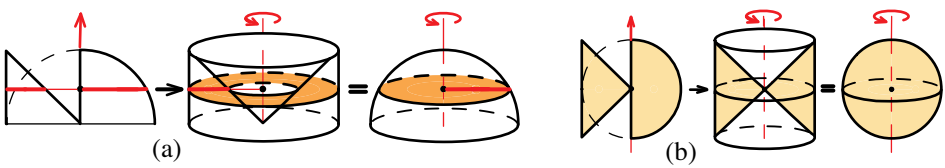


Figure 13.2: (a) Rotation of balanced chords produces cross sections of equal areas of a punctured cylinder and its insphere. (b) Rotation of balanced areas produces a punctured cylinder and its insphere of equal volumes.

Second, rotation of the two balanced areas in Figure 13.1b generates a punctured cylinder and the inscribed sphere as indicated in Figure 13.2b, which, by another theorem of Pappus, have equal volumes. This gives a second proof of Archimedes' volume result on the sphere and cylinder.

In [47; *Method*, p. 14] Archimedes makes the following comment: "*...I apprehend that some, either of my contemporaries or of my successors, will, by means of the method when once established, be able to discover other theorems in addition, which have not yet occurred to me.*"

This chapter does that. It introduces new ideas and methods that lead to many surprising results that Archimedes would have appreciated, all of which are extended to higher-dimensional space. One striking consequence is a natural extension to n-space of Archimedes' results on both the volume *and* surface area of a sphere, in which the cylinder is replaced by a "cylindroid," and the fraction $2/3$ is replaced by $2/n$ for all $n \geq 2$. The cylindroid is a cylinder only when $n = 3$.

Although the surface area relations for the sphere and cylinder do not follow from Archimedes' balancing in Figures 13.1a and 13.1b, they easily follow from our new balancing principles which imply, in particular, the arclength equilibrium in Figure 13.1c. This leads directly to a new proof of Archimedes' area result for the sphere and cylinder (see Figure 13.7b), and to a host of general surface area relations as well.

13.2 BALANCING REGULAR CIRCUMGONS IN A PLANE

Our methods are based on a new balancing lemma involving lengths of tangent segments to a circle, as depicted in Figure 13.3a. This simple lemma leads to profound consequences concerning equilibria of arclengths of curves and areas of plane regions, as well as surface areas and volumes of various solids.

Main balancing lemma.

Figure 13.3a shows a fixed vertical balancing axis, indicated by the arrow, passing through the center of a circle of radius r. Consider a tangent segment of given length L with its midpoint on the circle, to the right of the axis, and its projection, of length H, onto a vertical tangent line on the other side of the axis.

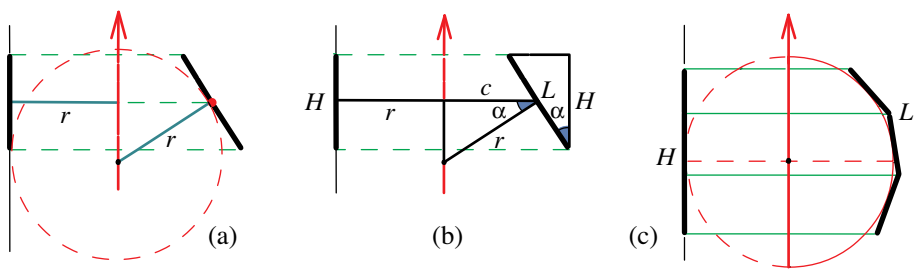


Figure 13.3: (a) Balancing a line segment and its projection. (b) Proof of the balancing principle. (c) Balancing a regular polygonal arc and its tangential projection.

If α is the angle between the tangent segment and the balancing axis, shown shaded in Figure 13.3b, we have $L\cos\alpha = H$. Multiplying by the radius r and introducing $c = r\cos\alpha$ we obtain the simple relation

$$Lc = Hr. \tag{13.1}$$

This equates moments of lengths L and H about the balancing axis, and establishes arclength equilibrium of the two segments: The length of a tangent segment (with

center on the circle) times the distance of its center from the balancing axis is equal to the length of its projection times its distance from the axis. We state this as:

Balancing Lemma 1. *A line segment tangent at its midpoint to a given circle is in arclength equilibrium, relative to an axis through the center of the circle, with its projection on a tangent line parallel to and on the other side of that axis.*

We turn next to extensions and applications of the balancing lemma.

Balancing regular circumgons with their tangential projections.

Circumgons are general objects circumscribing circles as introduced in Chapter 4. This chapter treats only *regular circumgons*, defined as follows: A regular circumgon is composed of several connected segments of equal length tangent at their midpoints to the incircle. Figure 13.3c shows an example on one side of a diameter used as a balancing axis, with corresponding projected segments shown on a common vertical tangent to the left of the axis. Because the length of each edge of the circumgon is in equilibrium with that of its projection as indicated in (13.1), the entire circumgon is in arclength equilibrium with its entire projection, which we refer to as the *tangential projection*. In other words, (13.1) is still valid, where now L denotes the total length of the circumgon, c denotes the distance of its centroid from the balancing axis, and H denotes the total length of its tangential projection. The midpoint requirement in Lemma 1 restricts our circumgons to be regular. Therefore we have:

Proposition 1. *A regular circumgon is in arclength equilibrium with its tangential projection.*

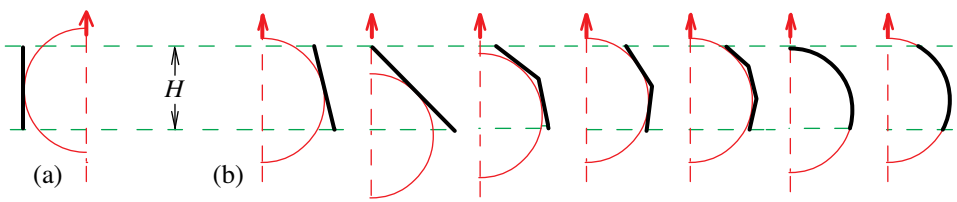


Figure 13.4: Balancing regular circumgons of a given height H in (b) with their tangential projections of total length H in (a). The circular arcs are limiting cases.

Figure 13.4b shows examples of regular circumgons of equal vertical height H in arclength equilibrium with a projected vertical segment in Figure 13.4a of the same height. All the circumgons have the same inradius but are located differently with respect to their common incircle. The last two examples are circular arcs, limiting cases of regular circumgons, whose arclengths are also in equilibrium with the same vertical projected segment of length H. Another example is the semicircular arc and its tangential projection in Figure 13.1c.

Balancing regular circumgonal regions with their tangential projection regions.

Figure 13.5 shows concentric regular circumgons in arclength equilibrium with their corresponding projections. By considering the union of an unlimited number of concentric circumgons in equilibrium with their corresponding projections, as suggested

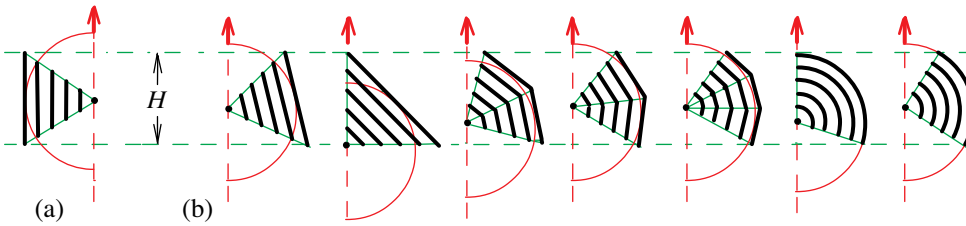


Figure 13.5: Balancing concentric regular circumgons in (b) with their respective projections in (a). The concentric circular arcs are limiting cases.

in Figure 13.6a, we can fill two-dimensional projection regions to also obtain areas in equilibrium as depicted by the examples in Figures 13.6b and c. Figure 13.6d shows the limiting case when the circumgons become circular arcs. This gives us:

Proposition 2. *A regular circumgonal region is in area equilibrium with its tangential projection region.*

The result can be written as an equation involving moments:

$$Ac_A = Pc_P, \tag{13.2}$$

where A is the area of the circumgonal region, P is the area of its projection region, and c_A, c_P are the respective centroidal distances from the balancing axis. This formula will be used later in Section 13.9.

In the special limiting case when the circumgonal region is a semicircular sector, the equilibrium in Figure 13.6d becomes Archimedes' area equilibrium in Figure 13.1b.

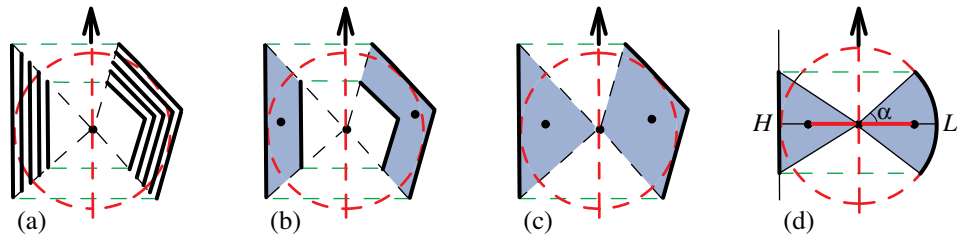


Figure 13.6: Circumgonal regions in area equilibrium with their corresponding tangential projection regions. The circular sector in (d) is a limiting case. The area equilibrium in (b) can also be verified directly by using known formulas for the centroid of a trapezoidal region.

Double equilibrium of regular circumgons.

Our process of building two-dimensional circumgonal regions as unions of concentric circumgons is somewhat analogous to Archimedes' process of building the two-dimensional triangle and semicircular disk in Figure 13.1b as unions of horizontal

chords in Figure 13.1a. At each stage, balancing of lengths leads to corresponding balancing of areas. But our process also preserves balancing of circumgonal arclengths, including the limiting circular arcs, which Archimedes' process does not. Thus, Propositions 1 and 2 together give us *double equilibrium*, which can be described as follows:

Proposition 3. *A regular circumgonal region, including the limiting case of a circular sector, is in area equilibrium with its tangential projection region; in addition, its outer boundary is in arclength equilibrium with its tangential projection.*

13.3 BALANCE-REVOLUTION PRINCIPLE AND CIRCUMSOLIDS

This section introduces a balance-revolution principle that enables us to determine lateral surface areas and volumes of circumsolids of revolution by reducing them to those of cylinders.

Lateral surface areas.

Balance-revolution principle for surface areas. *If two plane curves are in arclength equilibrium with respect to a balancing axis, then the surfaces of revolution generated by rotating them about the balancing axis have equal areas.*

Arclength equilibrium means that corresponding moments are equal, and the balance-revolution principle follows by multiplying each member of this equality by 2π and applying Pappus' rule which, for surfaces of revolution, states that:

The area of a surface of revolution swept by a plane curve is equal to the length of the curve times the circumference of the circle traced by the centroid of the curve.

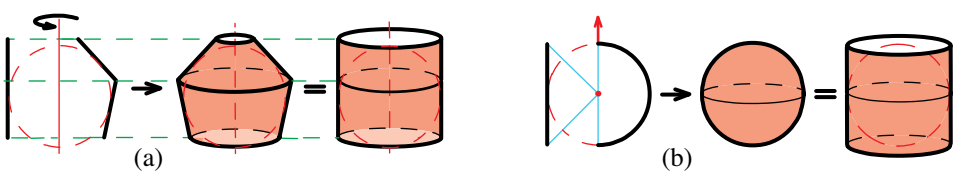


Figure 13.7: (a) Circumgon and its projection rotated about their balancing axis form surfaces of equal area. (b) Semicircle and its balanced projection sweep out sphere and cylinder of equal area, a new proof of Archimedes' result.

Figure 13.7a shows an example of a regular circumgon with two edges and its projection rotated together around their balancing axis. The rotated circumgon generates lateral surfaces of different truncated circular cones tangent to the insphere, and the vertical tangential projection generates a lateral cylindrical surface tangent to the insphere. By the balance-revolution principle, the area of the lateral surface of each truncated cone is equal to the lateral area of the corresponding *circumcylinder*. By additivity of area, the same is true for the composite surfaces. Although the circumgon in Figure 13.7a has only two edges, the same argument

works for any regular circumgon. Figure 13.7b proves Archimedes' result that a sphere and its circumcylinder have equal areas because they are swept by rotating a semicircular arc and its projection about their balancing axis. These examples illustrate the following:

Theorem 13.1. *The lateral surface area of revolution generated by a regular circumgon, including the limiting case of a circular arc, is equal to that of the circumcylinder whose height is that of the circumgon.*

In other words,

$$S = 2\pi r H, \tag{13.3}$$

where S is the lateral surface area generated by a regular circumgon of revolution with inradius r and total height H in the direction parallel to the axis of rotation. This striking result is exhibited in Figure 13.8, which shows surfaces of revolution

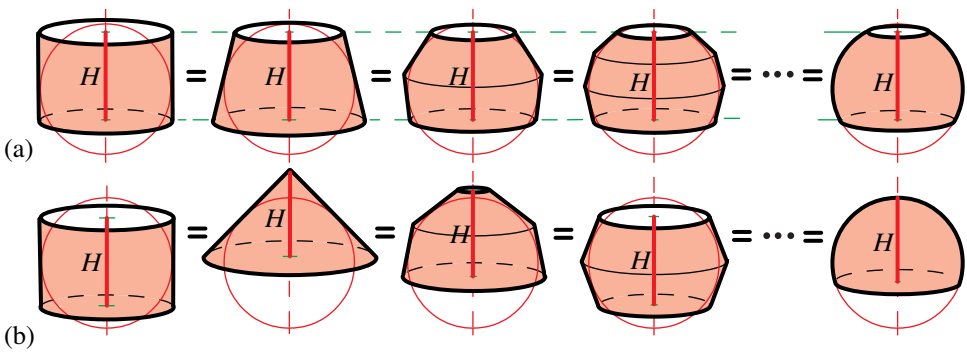


Figure 13.8: All regular circumgons of height H sweep out surfaces of area $2\pi r H$. All spherical zones of height H also have area $2\pi r H$.

generated by various circumgons of height H and the same inradius r. The same holds for the limiting case when the regular circumgon becomes a circular arc and its surface of revolution is part of a sphere. The area of a spherical zone of radius r lying between two parallel planes at distance H apart is $2\pi r H$, regardless of the location of the planes. Figure 13.8b shows circumgons located differently with respect to the diameter of the incircle. In particular, the surface area of a cone, truncated or not, with generating segments touching the insphere at their midpoints, is equal to that of a cylinder of the same height and the same insphere.

Consider the special case of Theorem 13.1 in which the circumgon is half a regular $2n$-gon rotated around a diameter of the incircle perpendicular to two opposite sides of the circumgon, as shown by the examples in Figure 13.9. We call these *right* circumgonal surfaces of revolution. They all have height $2r$. The circumscribing cylinder is the case $n = 2$. All the lateral surfaces have equal area $4\pi r^2$, which is also the area of the limiting sphere. This gives us the following remarkable extension of Archimedes' area result on the sphere and cylinder:

Corollary of Theorem 13.1. *All right circumgonal surfaces of revolution with inradius r have height $2r$ and lateral surface area $4\pi r^2$.*

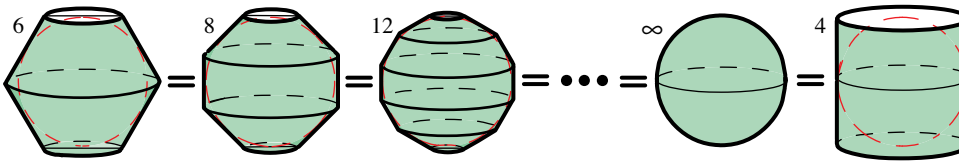


Figure 13.9: All right circumgonal surfaces of revolution of height $2r$, including the insphere, have equal lateral area, which is that of the circumcylinder, $4\pi r^2$.

Volumes.

Figure 13.10b shows two circumgonal regions in area equilibrium obtained by balancing, as was done earlier in Figure 13.6c, a collection of concentric circumgons and their corresponding projections in Figure 13.10a. When these regions are rotated

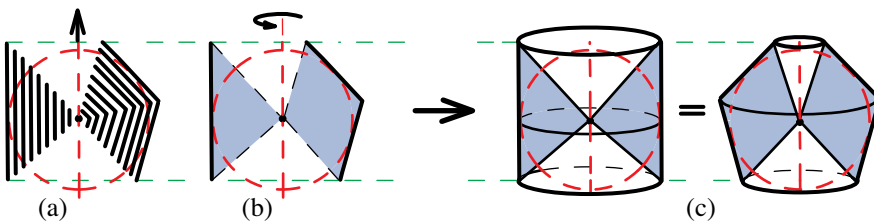


Figure 13.10: Rotating two circumgonal regions in area equilibrium in (b) gives two punctured solids of revolution in (c) with equal volumes.

together about the balancing axis they sweep out two solids of revolution having equal volumes because of the following principle, illustrated in Figure 13.11.

Balance-revolution principle for volumes. *If the areas of two plane regions are in equilibrium with respect to a balancing axis, then the solids of revolution generated by rotating them about the balancing axis have equal volumes.*

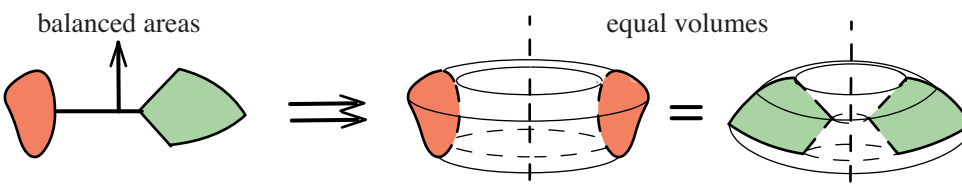


Figure 13.11: Rotating two plane regions in area equilibrium about the balancing axis produces solids of revolution having equal volumes.

The balance-revolution principle for volumes follows from Pappus' rule for volumes, which is analogous to that for surfaces, applied to solids of revolution.

The two solids shown in Figure 13.10c have equal volumes. They are obtained

by rotating the triangle and quadrilateral in area equilibrium in Figure 13.10b. One is a solid circular cylinder punctured by two cones, and the other consists of two parts, each being a frustum of a solid cone punctured by another cone.

Figure 13.12 shows further examples illustrating the balance-revolution principle. Each pair of solids is generated by rotating about the balancing axis two circumgonal regions in area equilibrium, so they have equal volumes.

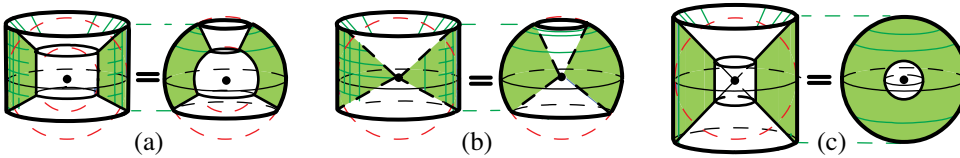


Figure 13.12: Circumsolids of revolution. Equal heights and volumes in each pair.

Figure 13.13 shows solids of revolution generated by regular circumgonal plane regions and their tangential projections, all of altitude H. Each solid is punctured by two cones whose vertices are at the center of the insphere. Because they circumscribe

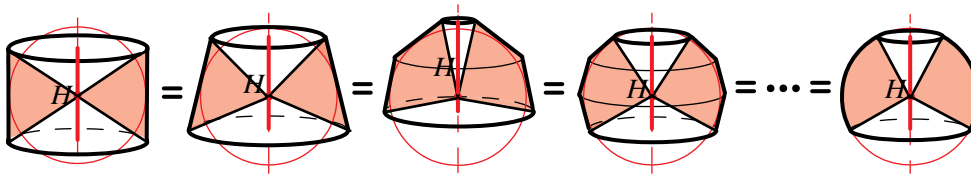


Figure 13.13: Regular circumgonal regions of equal altitude in area equilibrium generate punctured solids of revolution of equal volume.

a sphere, all are examples of *circumsolids*, as described in Chapter 4, and they illustrate the following general theorem:

Theorem 13.2. *The volume of a punctured circumsolid generated by rotating a regular circumgonal region is equal to that of the punctured circumcylinder whose height is that of the circumgon.*

Each such solid has volume V of a punctured cylinder given by

$$V = \frac{2}{3}\pi r^2 H. \tag{13.4}$$

This result is consistent with our knowledge of circumsolids. Each solid of revolution is a circumsolid so, by Theorem 4.11, each volume is one-third the product of its outer surface area, $2\pi rH$, and the radius of the insphere, in agreement with (13.4).

Figure 13.14 shows examples of *right circumsolids*, obtained by rotating half a regular $2n$-gonal region of the type shown in Figure 13.9. For such circumsolids we have the following extension of Archimedes' volume result on a sphere and its punctured circumscribing cylinder, which is the case $n = 2$:

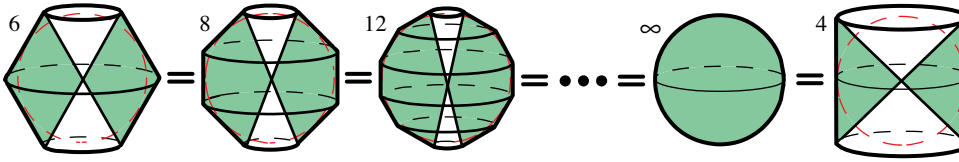


Figure 13.14: All punctured right circumgonal solids of revolution of height $2r$ have the same volume as the insphere and its punctured circumscribing cylinder, $4\pi r^3/3$.

Corollary of Theorem 13.2. *All punctured right circumsolids of revolution with inradius r have height $2r$ and volume $4\pi r^3/3$.*

13.4 MOMENT-WEDGE PRINCIPLE AND CYLINDRICAL WEDGES

As noted, Archimedes used area balance in Figure 13.1b to find the volume of a cylindrical wedge, the main new result of his Method. Using a new principle, we shall obtain volumes of a family of wedges with circumgonal bases, which includes the Archimedes wedge as a limiting case. A similar treatment is given for their lateral surface areas. Incidentally, Archimedes did not treat the surface area of a cylindrical wedge.

Moment-wedge volume principle and circumgonal wedges.

This principle, stated in (13.5), relates the area moment of the planar base of a right cylinder, with respect to an axis in the plane of the base, with the volume of a wedge cut from the cylinder by an inclined plane. Figure 13.15a shows a right cylinder with a horizontal planar base of general shape cut by a plane whose angle of inclination with the horizontal has tangent k. We are interested in the area moment of the base relative to the line of intersection of the two planes, which we take as an axis of moments. An element of area $\Delta x\Delta y$ at distance x from the axis has moment $\Delta M = x\Delta x\Delta y$. On the other hand, the cylindrical column shown is an element of volume given by $\Delta W = kx\Delta x\Delta y$. Therefore,

$$\Delta W = k\Delta M.$$

Integration gives

$$W = kM, \tag{13.5}$$

where W is the volume of the cylindrical wedge whose base has area moment M about the horizontal y axis.

In Figure 13.15b we apply (13.5) to two cylindrical wedges whose horizontal bases are the triangle and semicircular disk of Figure 13.1b. As we know, the two bases are in area equilibrium about the y axis, so moment M is the same for them if the two cutting planes are inclined at the same angle with the horizontal. Hence, by (13.5), the two wedges have equal volumes $W = kM$. Thus, the volume of the cylindrical wedge is equal to that of the pyramid whose rectangular base is

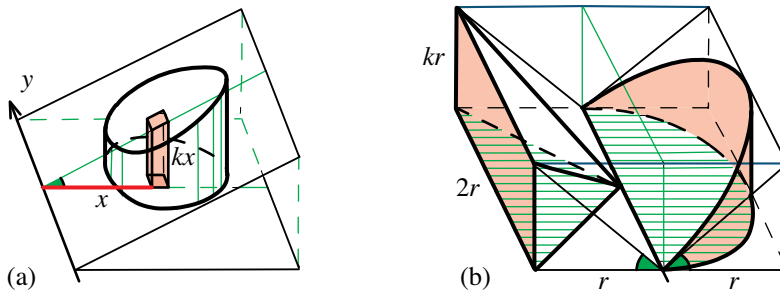


Figure 13.15: (a) Relating the volume of a general cylindrical wedge to the area moment of its base. (b) Volume of Archimedes' cylindrical wedge is equal to that of the pyramid because their bases are in area equilibrium.

shaded in Figure 13.15b. This gives $W = 2kr^3/3$, one-third the area of the base times the altitude. Archimedes deduced (by a completely different method) the equivalent result that the volume of the cylindrical wedge is one-sixth that of the largest rectangular box in Figure 13.15b.

The same idea applied to two general cylindrical wedges whose bases are in area balance gives the following corollary:

Balance-wedge volume principle. *Two cylindrical wedges have equal volumes if their bases are in area equilibrium about the line of intersection of the two cutting planes inclined at the same angle with the common base plane.*

Right cylinders include prisms with circumgonal bases. We apply the corollary to prisms with circumgonal bases in area balance. Those in Figure 13.16 are built from halves of regular $2n$-gons in area balance with the same triangular base. All the wedges have volume equal to that of the pyramid with rectangular base.

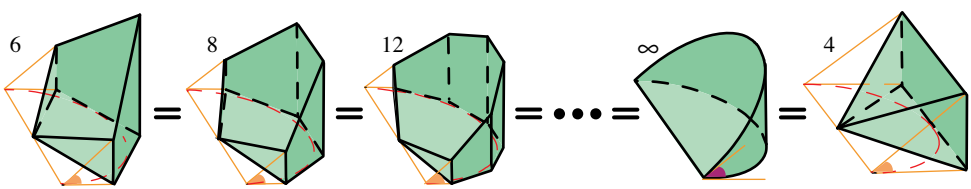


Figure 13.16: Circumgonal wedges of equal volume. Their bases are halves of regular $2n$-gons, all in area balance. Archimedes' wedge in Figure 13.15b is their limiting case.

Lateral surface area of a circumgonal wedge.

If the circle in Figure 13.15b has radius r, the shaded rectangle has base $2r$, altitude kr, and area $2r^2k$. This is also the lateral surface area of the cylindrical wedge, shaded in Figure 13.15b. To verify this we note that by double equilibrium

(Proposition 3), the semicircle and its horizontal projection in the plane of the base are in arclength equilibrium. Thus an element of arc of length Δs and its projection Δp have equal moments about the diameter, $\Delta s \cdot x = \Delta p \cdot r$, where x is the distance of element Δs from the diameter. The corresponding elements of lateral surface area are $\Delta s \cdot kx$ for the cylinder and $\Delta p \cdot kr$ for the projection, which are equal because the moments are equal. The sum of the Δp is $2r$ so by integration with respect to arclength we see that the lateral surface of the cylindrical wedge is $2r^2k$, the same as that of the projection rectangle. The same applies to the lateral surface of each wedge in Figure 13.16, which consists of trapezoidal faces whose total area is equal to that of the common projection rectangle. A similar proof works for a general cylindrical wedge like that in Figure 13.15a whose planar base is bounded by a rectifiable curve.

Thus, we have shown that the volume W of a cylindrical wedge and its lateral surface area A are given by

$$W = \frac{2}{3}kr^3, \ A = 2kr^2.$$

The results in this section will be extended to higher dimensions in Section 13.8.

Figure 13.17 shows examples of those lateral surfaces in Figure 13.16 with $r = k = 1$. Only half the unwrapped lateral surface is shown, the other half being symmetric. Each unwrapped surface shown in Figure 13.17 has area 1, which is the area of the common projection unit square. In the limiting case when the base is circular, the unwrapped portion of the cylindrical surface is the region under half a sine curve.

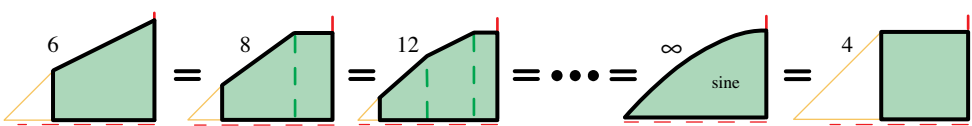


Figure 13.17: Halves of unwrapped lateral surfaces of the circumgonal wedges in Figure 13.16. All have the area of a unit square, the same as the region under half a sine curve.

13.5 BALANCING PORTIONS OF A SPHERE AND OF A CYLINDER

This section extends Balancing Lemma 1 by establishing equilibria of various portions of a sphere and of a circular cylinder with their tangential projections.

Surfaces on a sphere.

We can generate surfaces on a sphere by rotating a circular arc and its projection about a horizontal axis through the center of the circle, as indicated in Figure 13.18a. Let Δs denote the length of the small circular arc in Figure 13.18a. Then its projection has length Δp which satisfies the approximate relation $\Delta s \cdot \cos\alpha = \Delta p$,

where α is the angle of inclination shown. Multiplying by the radius r to obtain moments about the balancing line, we find

$$\Delta s \cdot (r \cos \alpha) = \Delta p \cdot r,$$

which can be regarded as an approximate balancing relation. Rotating the balancing line generates a central balancing plane, and rotating the projection line generates a projection plane, both perpendicular to the axis of rotation, as in Figure 13.18b. The circular arc generates an element of spherical surface area ΔS obtained by

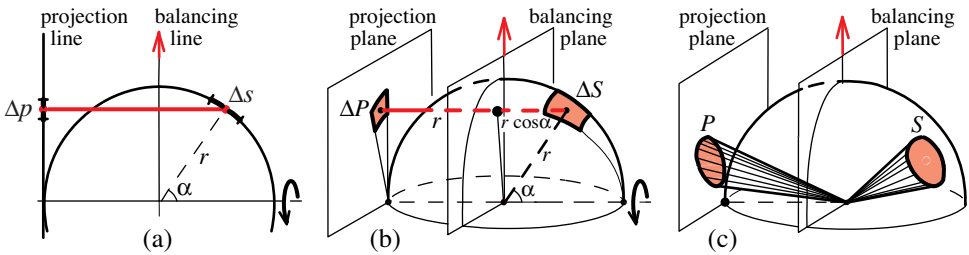


Figure 13.18: Rotating a circular arc and its vertical projection in (a) about a horizontal axis generates a surface element in (b) in area equilibrium with its projection relative to a balancing plane. (c) Solid angle in double equilibrium with its projection cone.

rotation through a small angle. Let ΔP denote the area of the projection of ΔS on the tangent projection plane. They are in the approximate relation $\Delta S \cdot \cos \alpha = \Delta P$. Multiplying by r to obtain moments, we find

$$\Delta S \cdot (r \cos \alpha) = \Delta P \cdot r, \tag{13.6}$$

which is an approximate area balancing relation. The factor $r \cos \alpha$ is the centroidal distance of a small spherical surface area element from the balancing plane, r is the centroidal distance of the projection from the balancing plane, and (13.6) represents area equilibrium of elements ΔS and ΔP with respect to this plane.

In particular, each small area element in Figure 13.18b will be in area equilibrium with its corresponding planar projection according to (13.6). Because any region of area S on the sphere and its projection of area P on the projection plane can be approximated arbitrarily closely by such area elements, (13.6) gives us the area equilibrium relation

$$Sc = Pr, \tag{13.7}$$

where r is the radius of the sphere and c is the centroidal distance of the surface of area S from the balancing plane. This relation, illustrated in Figure 13.19, can be stated as follows:

Balancing Lemma 2. *Any portion of a spherical surface is in area equilibrium with its tangential projection relative to a balancing plane through the center of the sphere.*

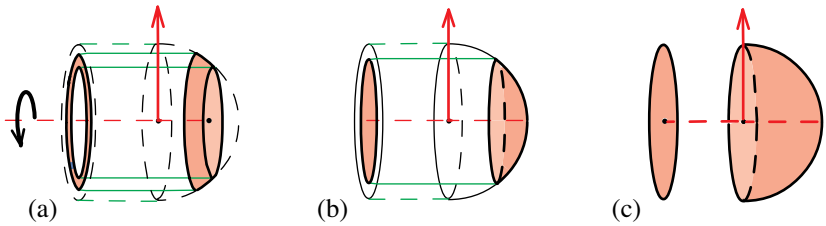


Figure 13.19: Balancing: (a) spherical zone and its projection annulus; (b) spherical cap and its projection disk; (c) hemispherical surface and its projection disk.

Double equilibrium of a solid angle and its projection cone.

Figure 13.18c shows a solid angle of radius r, a union of radial lines emanating from the center of the sphere to a portion of the sphere with area S. It is also a union of concentric layers of spherical surfaces formed by radial shrinking of the spherical surface of area S to 0. The spherical surface of each layer is in area equilibrium with its projection on a plane tangent to the concentric sphere. By multiplying (13.6) by the thickness Δr of a typical layer, we obtain layer by layer volume equilibrium. Consequently, the entire solid angle is in volume equilibrium with its projection solid, a cone whose vertex is at the center of the sphere and whose base is the projection of the region of area S. The solid spherical sector in Figure 13.20a is

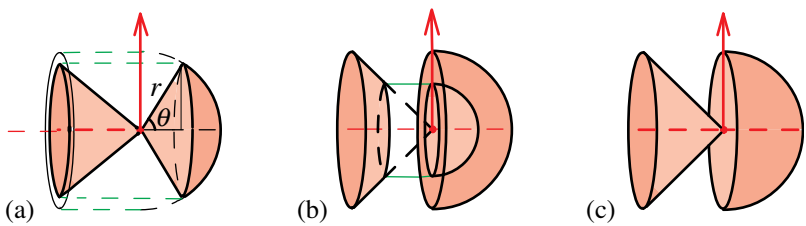


Figure 13.20: Double equilibrium of: (a) spherical sector and cone; (b) hemisphere with cavity and truncated cone; (c) hemisphere and solid cone.

a special case. It can also be formed from a collection of concentric hemispherical caps in area equilibrium with corresponding projected disks as in Figure 13.19b to produce a spherical sector in volume equilibrium with its projection cone.

Similarly, we can use concentric hemispherical surfaces like those in Figure 13.19c to build a solid hemisphere in Figure 13.20c in volume equilibrium with a solid cone built from the corresponding projected equatorial disks. Figure 13.20b shows an intermediate stage, a solid hemisphere with cavity balanced by a truncated cone.

Balancing axes for symmetric objects.

The equilibria in Figures 13.19 and 13.20 are with respect to a balancing plane. All the objects were obtained by rotation about a horizontal axis, which is also an axis

of symmetry that intersects the balancing plane at the centroid of each composite object. Consequently, any axis through the centroid will also be a balancing axis. This also applies to more general situations when the centroid of a composite object is determined by symmetry, as in Figures 13.1, 13.19, 13.20, 13.21a, 13.22c, 13.23, 13.24, 13.27, 13.29a, and 13.29c.

Double equilibrium: spherical wedge and elliptical cone; punctured spherical zone and projection cone.

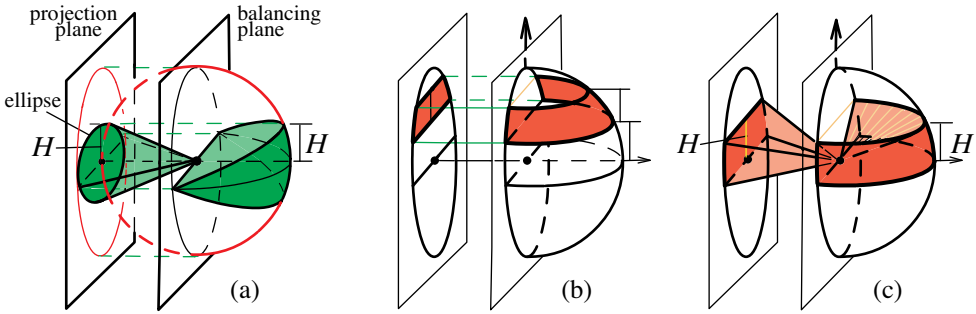


Figure 13.21: (a) Double equilibrium: spherical wedge and cone. (b) Area equilibrium: zone and its projection. (c) Double equilibrium: punctured spherical zone and cone.

Figure 13.21a shows a spherical wedge of height $2H$ cut from a solid hemisphere of radius r by two planes through a diameter, each making an angle θ with the equatorial plane. The spherical surface of the wedge is in area equilibrium with its projection, which is an ellipse whose major axis has length $2r$ and whose minor axis has length $H = 2r \sin\theta$. The solid spherical wedge is also in volume equilibrium with the solid elliptical cone. Figure 13.21b shows a spherical surface of a zone in area equilibrium with a slice of a circular disk, and Figure 13.21c shows double equilibrium of solids built from concentric figures of the type in Figure 13.21b.

Balancing portions of a circular cylinder.

Instead of rotating the circle in Figure 13.18a to generate a sphere, we can translate it in a direction perpendicular to its plane to generate a circular cylinder of radius r, as suggested by Figure 13.22a. Corresponding translations of the balancing line and projection line in Figure 13.18a generate two parallel planes that we call the balancing plane (a bisector of the cylinder) and the projection plane (tangent to the cylinder), as shown in Figure 13.22a. Translation of the circular arc in Figure 13.18a sweeps out an area element (on the lateral surface of the cylinder) in area balance with its vertical projection rectangle (on the projection plane), as indicated in Figure 13.22a. More generally, any region of area S on the cylinder will be in area equilibrium, with respect to the balancing plane, with its corresponding projection of area P on the projection plane, as illustrated in Figure 13.22b, because they can

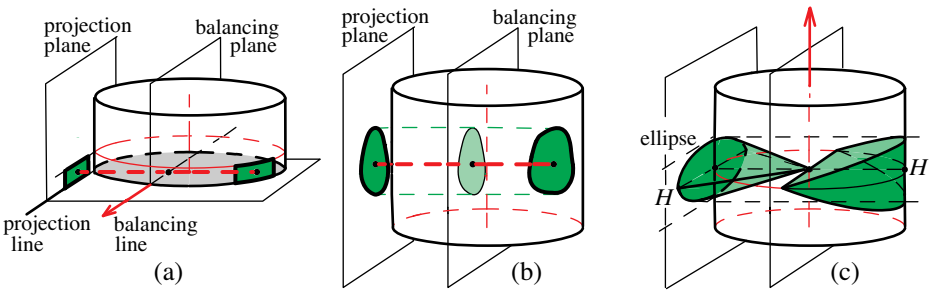


Figure 13.22: (a) Small area element on cylinder in area balance with rectangular projection. (b) Region on cylinder in area balance with tangential projection. (c) Cylindrical wedge in double equilibrium with its projection elliptic cone.

be approximated arbitrarily closely by area elements like those in Figure 13.22a. Again, (13.7) holds, where now S is the area of the cylindrical region, P is the area of its projection, and c is the centroidal distance from the balancing plane to the cylindrical surface of area S. This gives us:

Balancing Lemma 3. *Any portion of a cylindrical surface is in area equilibrium with its tangential projection relative to a balancing plane through the axis of the cylinder.*

Figure 13.22c shows a solid cylindrical wedge of height H between two inclined planes, each making an angle θ with the horizontal equatorial plane. The lateral surface of the wedge is in area equilibrium with its projection, which is an ellipse whose major axis has length $2r$ and whose minor axis has length H. By using concentric cylindrical wedges of decreasing radii, we can build a solid cylindrical wedge in double equilibrium with a solid elliptic cone as shown in Figure 13.22c. Not only is the lateral cylindrical surface in area equilibrium with its elliptical projection, but also the solid cylindrical wedge is in volume equilibrium with the elliptic cone.

The spherical wedge in Figure 13.21a and the cylindrical wedge in Figure 13.22c are in double equilibrium with the same elliptic cone. Eliminating the cone, we obtain the double equilibrium in Figure 13.23, which can be described as follows:

Proposition 2. *A spherical wedge is in double equilibrium with a cylindrical wedge of the same height and the same insphere.*

When the wedges in Figure 13.23 are viewed along the common diameter, we see the area and arclength equilibrium obtained earlier in Figure 13.6d.

Figure 13.24a shows Figure 13.21a when the wedge is a hemisphere, and Figure 13.24b shows the cylindrical wedge and cone in Figure 13.22c when the altitude is a diameter. The equilibrium in Figure 13.24c is obtained by eliminating the cone in the previous two figures. Figure 13.24d shows a triangular prism and semicircular cylinder also in double equilibrium. Archimedes established their volume equilibrium, which is equivalent to area equilibrium of the triangle and semicircular disk

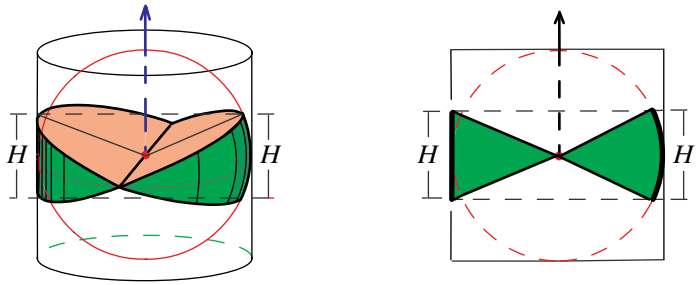


Figure 13.23: Spherical and cylindrical wedges in double equilibrium.

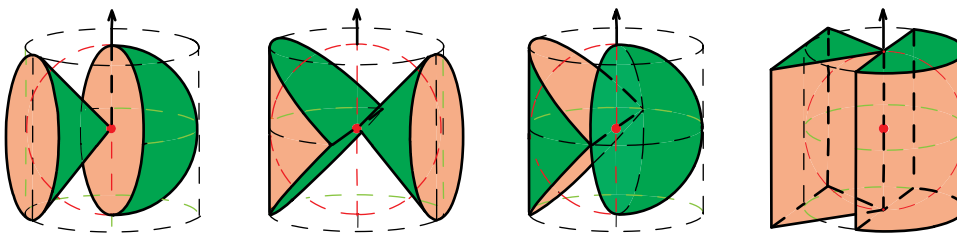


Figure 13.24: Solids in double equilibrium. The top view of (d) appears in Figure 13.1b; it also represents the top view of each of (a) and (b) and the side view of (c).

in the top view.

The balancing relations just established, although of interest in their own right, will be extended in the next section and applied to obtain equality of volumes and surface areas of higher-dimensional solids.

13.6 HIGHER-DIMENSIONAL BALANCING PRINCIPLES

We begin this section by analogy with the process in Figure 13.2b, where the balanced 2-dimensional regions were rotated in 3-space about the balancing axis to generate a solid sphere and punctured cylinder. Now we do the same with the cone and hemisphere in equilibrium in Figure 13.24a. Imagine these solids imbedded in 4-space and rotate each around the balancing axis to sweep out 4-dimensional solids. The hemisphere generates a 4-dimensional solid sphere, and the cone generates an object which we call a *punctured* 4*-cylindroid*. To find relations between their volumes and surface areas we need a higher-dimensional balance-revolution principle.

Double equilibrium of solid angle with its projection cone.

We will extend some of our earlier balancing relations to higher-dimensional space in a manner completely analogous to the transition from 2 dimensions in Figure 13.18a

to 3 dimensions in Figure 13.18b. First, regard the 3-dimensional configuration in Figure 13.18b as imbedded in 4-space, and rotate the configuration in 4-space around the horizontal axis of symmetry of the hemisphere. The 3-dimensional hemisphere to the right of the balancing plane produces a 4-dimensional hemisphere, the balancing plane produces a balancing hyperplane, and the projection plane produces a projection hyperplane in 4-space. If rotated through a small angle, the 3-dimensional surface element produces a corresponding surface element on the 4-sphere, and the surface element on the projection plane produces a surface element of the projection hyperplane in 4-space, which is the projection of the surface element on the 4-sphere.

The same type of argument that produced the balancing relations (13.6) and (13.7) shows that two elements in 4-space satisfy the same kind of balancing relation with respect to the balancing hyperplane.

By repeating the process and arguing by induction on the dimensionality of the space, we find that the same type of balancing holds in n-space, because the angle of inclination α of the radial line being rotated in Figures 13.18a and 13.18b is the same for all dimensions. This leads to double equilibrium, with respect to the balancing plane, of an n-dimensional solid angle and its projection n-cone built from parallel projection bases, by analogy with the solid angle equilibrium described in Figure 13.18c.

Solid angle balancing lemma. *An n-dimensional solid angle and its projection n-cone are in double equilibrium with respect to a balancing hyperplane that passes through the vertex of the cone and is parallel to the projection hyperplane.*

Moment-volume principle.

The following principle relates moments and "volumes" of revolution. We place quotation marks around the word "volume" to state one principle that applies to both surface areas and volumes as understood in the traditional use of the terms.

Moment-volume principle. *When an object in n-space with moment M_n relative to an axis is rotated in $(n+1)$-space about that axis, it produces an object whose "volume" V_{n+1} is given by*

$$V_{n+1} = 2\pi M_n. \tag{13.8}$$

In other words, the volume of an $(n+1)$-dimensional solid of revolution is 2π times the n-dimensional moment relative to the axis of rotation of the object that generates the solid.

The idea of the proof can be easily seen when $n = 2$. Imagine an area element $\Delta x \Delta y$ at distance x from the y-axis in the xy plane. Its moment about the y axis is $\Delta M_2 = x\Delta x\Delta y$. When this is rotated about the y-axis it sweeps out a solid of volume $\Delta V_3 = 2\pi x\Delta x\Delta y = 2\pi\Delta M_2$. Similarly, for higher dimensions we find

$$\Delta V_{n+1} = 2\pi\Delta M_n.$$

Integration then gives (13.8).

In particular, we have the following corollary:

Balance-revolution principle. *If two objects in n-space are in "volume" equilibrium with respect to a balancing axis, then the objects in $(n+1)$-space generated by rotating them about this axis have equal "volumes."*

In the balance-revolution principle for surface areas stated at the beginning of Section 13.3, the objects in 2-space in volume equilibrium are two plane curves in arclength equilibrium. When they are rotated about the balancing axis they generate surfaces in 3-space with equal volumes, which now means equal surface areas. In the balance-revolution principle for volumes stated later in Section 13.3, the objects in 2-space in volume equilibrium are two plane regions in area equilibrium. When they are rotated about the balancing axis they generate solids in 3-space with equal volumes, which now means ordinary volumes.

Applications to n-hemispheres.

The following specialization of the solid angle balancing lemma plays a key role in the sequel:

Special balancing lemma. *An n-hemisphere is in double equilibrium with its projection n-cone with respect to the vertex of the cone and to any axis through it.*

This special lemma, combined with the balance-revolution principle, leads to:

Double equality principle. *The objects obtained by rotating the n-hemisphere and n-cone in $(n+1)$-space have equal lateral surface areas and equal volumes.*

13.7 ON THE SPHERE AND CYLINDROID IN n-SPACE

The main result of this section is an extension of Archimedes' results on volumes and surface areas of spheres to n-space for all $n \geq 2$, stated below in Theorem 13.3. This is perhaps the most profound consequence of our balancing methods because when $n \neq 3$ no simple relation connects the volume or the surface area of an n-sphere with that of its circumscribing n-cylinder. (See [52] and [64] for attempts to find such a relation.) In our extension process a new object occurs naturally, an n-cylindroid, which has a simple and direct relation to its insphere. For $n = 3$ a cylindroid is a traditional cylinder (see Figure 13.26), but when $n \neq 3$ it is an entirely different object (see Figure 13.25 for the case $n = 2$).

Next we introduce a general cylindroid and compare with the standard cylinder. A classical n-cylinder is produced by translating an $(n-1)$-sphere in n-space. To produce an n-cylindroid, we tumble (rotate) an $(n-1)$-cylinder in n-space about one of its bases. To elaborate on the constructions, we describe both the cylinder and the cylindroid in successively higher-dimensional spaces, starting with $n = 2$, shown in Figure 13.25.

In Figure 13.25a, a one-dimensional sphere (a line segment) is translated in a perpendicular direction to produce a 2-cylinder (a rectangle). Rotating (or tumbling) a 1-cylinder (a line segment) about one end as in Figure 13.25b produces a 2-cylindroid (which is also a circular disk).

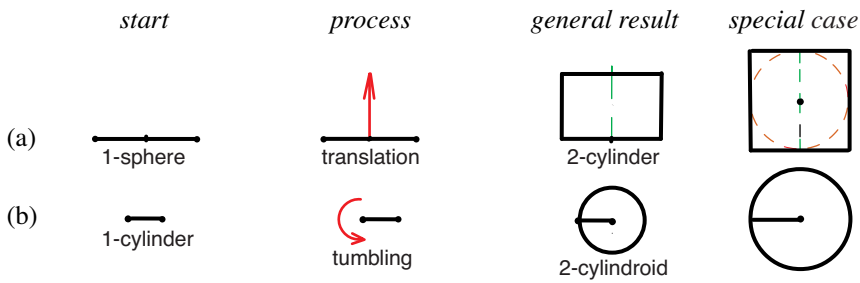


Figure 13.25: (a) Generating a 2-cylinder by translating a 1-sphere (line segment) in 2-space. (b) Generating a 2-cylindroid by tumbling (rotating) a 1-cylinder (line segment) about one end in 2-space. The 2-cylindroid is also a circular disk.

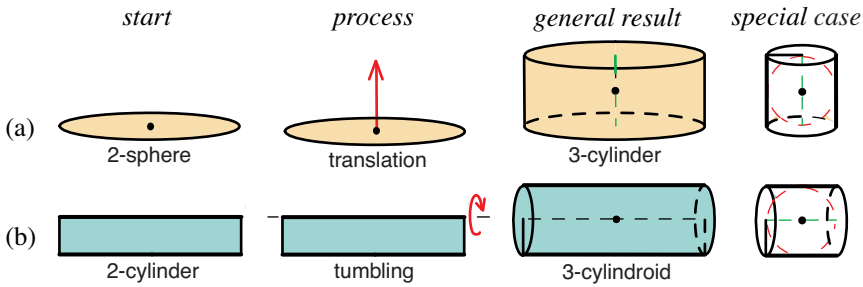


Figure 13.26: (a) Generating a 3-cylinder by translating a 2-sphere (disk) in 3-space. (b) Generating a 3-cylindroid by tumbling a 2-cylinder (a rectangle) in 3-space. The 3-cylindroid is also a 3-cylinder.

In Figure 13.26a, a 2-sphere (a circular disk) is translated in a perpendicular direction to produce a 3-cylinder, and in Figure 13.26b, a 2-cylinder (a rectangle) is tumbled about one of its bases to produce a 3-cylindroid, which in this case is also a 3-cylinder. There is no 1-cylindroid because there is no 0-cylinder to tumble.

Natural evolution of the cylindroid.

Figures 13.27 and 13.28 illustrate how our balancing methods lead naturally to the concept of cylindroid. In Figure 13.27, at the top of the first column, the semicircular disk and triangle are in double equilibrium about a vertical axis through a diameter. When these are rotated in 3-space about the balancing axis they produce two solids of revolution, the sphere and punctured cylinder shown below them, which, by the double equality principle, have equal volumes and equal lateral surface areas. This is Archimedes' classical result for $n = 3$, which we deduced by using double equilibrium.

When the triangle and semicircular disk in the first row are rotated in 3-space

about the *horizontal* axis of symmetry they produce a solid hemisphere and cone, in double equilibrium about the same vertical axis, shown at the top of the second column in Figure 13.27. Rotating these solids, in turn, in 4-space about the balancing axis produces two 4-dimensional solids of revolution shown at the bottom of the second column, which, by the balance-revolution principle, have equal "volumes," which here means both 4-dimensional volumes and lateral surface areas are equal.

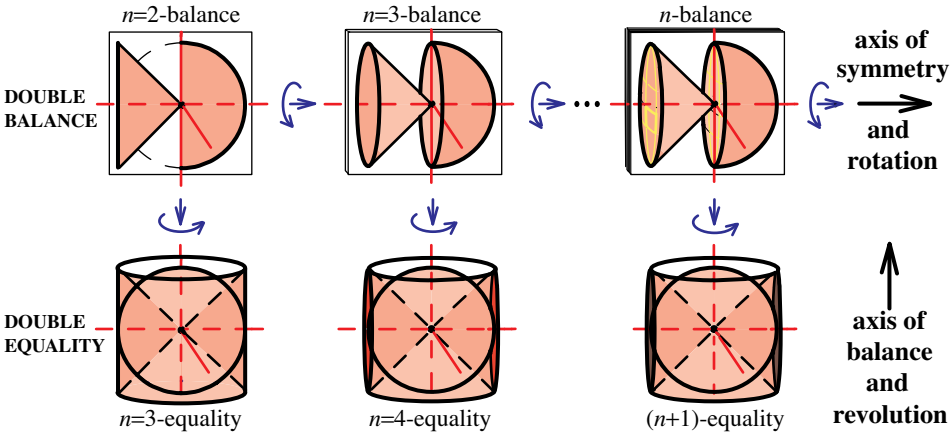


Figure 13.27: Diagram illustrating the natural evolution of the cylindroid.

To help visualize the process, refer to Figure 13.28a, showing a solid cone inscribed in a circular cylinder in 3-space. Figure 13.28b shows a schematic "flattening" of 3-space into a hyperplane in which the cone and cylinder appear as a triangle inscribed in a rectangle.

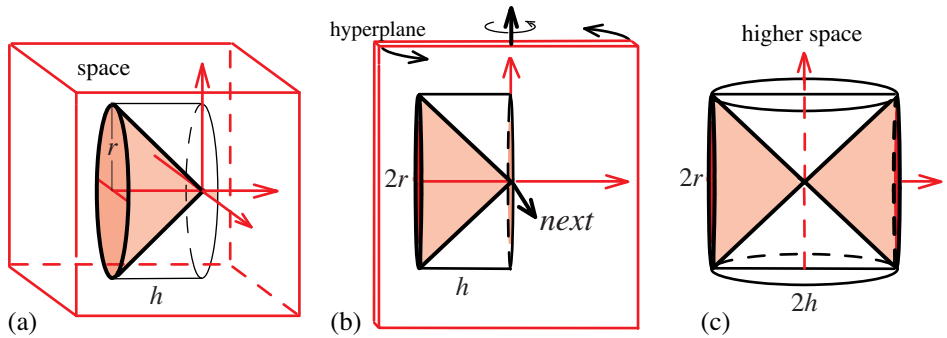


Figure 13.28: Constructing a 4-dimensional cylindroid and a punctured cylindroid. (a) Solid cylinder and inscribed cone in 3-space. (b) Schematic flattening of (a) onto a hyperplane, with a 4th axis erected perpendicular to it. (c) Rotation of (b) in 4-space around a coordinate axis perpendicular to the 4th axis and to the axis of symmetry. We refer to this as tumbling the figure.

We choose a 4th coordinate axis perpendicular to this hyperplane (labeled as *next* in Figure 13.28b) then rotate the hyperplane with its contents around an axis that is perpendicular to the 4th axis and also to the axis of symmetry of the cone and cylinder. We refer to this as *tumbling* the figure. If we rotated about the axis of symmetry, we would obtain a 4-cylinder with an inscribed 4-cone. Instead, by tumbling we obtain new objects, a 4-*cylindroid*, and the subset swept by the cone, a *punctured* 4-*cylindroid*, depicted in Figure 13.28c. The flattened base of the triangle in Figure 13.28b is actually the circular base of the cone in Figure 13.28a. Tumbling the base generates the lateral surface of the cylindroid in Figure 13.28c.

We continue the process in Figure 13.27, successively rotating the double-balanced objects in the top row through one higher dimension around the horizontal axis of symmetry to obtain a higher-dimensional hemisphere and cone in double equilibrium about the same vertical balancing axis through the vertex of the cone. When these balanced symmetric objects are tumbled through one higher dimension around the balancing axis they produce two solids of revolution of equal "volumes," a higher-dimensional sphere and a punctured higher-dimensional cylindroid, the key that enables us to extend Archimedes' results to n-space for every $n \geq 2$.

Definitions of cylindroid and punctured cylindroid.

In general, an n-cylindroid and a punctured n-cylindroid are defined by direct analogy with the 4-dimensional case. We begin with an $(n-1)$-cylinder of radius r and altitude h and the corresponding inscribed isosceles $(n-1)$-cone by analogy with Figure 13.28a. Tumble them together in n-space around a coordinate axis that passes through the vertex of the cone and is perpendicular to the axis of the cylinder. During the tumbling, the $(n-1)$-cylinder sweeps out an object we call an n-*cylindroid* of radius h and altitude $2r$. As the $(n-1)$-cylinder sweeps out the cylindroid, the $(n-1)$-cone sweeps out a portion of the n-cylindroid that we call a *punctured* n-*cylindroid*.

The common base of the $(n-1)$-cylinder and cone is an $(n-2)$-sphere of radius r. Cross-sections of the $(n-1)$-cylinder perpendicular to its axis are $(n-2)$-spheres of radius r, whereas cross-sections of the $(n-1)$-cone are $(n-2)$-spheres whose radii decrease linearly from r at the base to 0 at the vertex.

When $n \neq 3$, an n-cylindroid differs from an n-cylinder. In particular, when $n = 2$, a 2-dimensional cylinder is a rectangle, but a 2-dimensional cylindroid is a circular disk, obtained by tumbling a 1-cylinder (a line segment) about one of its endpoints. Moreover, the "cone" inscribed in the 1-cylinder is the 1-cylinder itself so the punctured and unpunctured 2-cylindroid are the same object.

Moment relations for cone and cylinder.

The following formulas for the volume $v_{n-1}(\text{cyl})$ of the $(n-1)$-cylinder and the volume $v_{n-1}(\text{cone})$ of an $(n-1)$-cone are known:

$$v_{n-1}(\text{cyl}) = hV_{n-2}, \quad v_{n-1}(\text{cone}) = \frac{v_{n-1}(\text{cyl})}{n-1} = \frac{h}{n-1}V_{n-2},$$

where V_{n-2} is the volume of their common $(n-2)$-spherical base, and h is their common altitude. It is also known that the centroid of the $(n-1)$-cylinder is at the midpoint of its altitude, while that of the $(n-1)$-cone is at a distance h/n from the base or, equivalently, at distance $h(n-1)/n$ from the vertex. Consequently, the volume moments of the cylinder and cone with respect to an axis through its vertex perpendicular to its axis of symmetry are given by

$$M_{n-1}(\text{cyl}) = \frac{h^2}{2}V_{n-2}, \quad M_{n-1}(\text{cone}) = \frac{h^2}{n}V_{n-2}. \tag{13.9}$$

From (13.9) we immediately obtain the following lemma that plays an important role in our extension of Archimedes' results.

Moment ratio lemma. *The volume moment of an $(n-1)$-cone with respect to an axis through its vertex is $2/n$ times the corresponding moment of its circumscribing $(n-1)$-cylinder; hence the ratio of the two moments, cone to cylinder, is $2/n$.*

Extension of Archimedes' classical results to n-space (suitable for engraving on Archimedes' hypertombstone).

Now we can state and prove the main result of this section, valid for all $n \geq 2$:

Theorem 13.3. (a) *The volume of an n-sphere equals that of its punctured circumscribing n-cylindroid.*

(b) *The surface area of an n-sphere is equal to the lateral surface area of its circumscribing n-cylindroid.*

(c) *The volume of an n-sphere is $2/n$ times that of the volume of its (unpunctured) circumscribing n-cylindroid.*

(d) *The surface area of an n-sphere is $2/n$ times that of the total surface area of its (unpunctured) circumscribing n-cylindroid.*

Proof of (a) *and* (b): These follow directly from the double equality principle stated at the end of Section 13.6.

Proof of (c) *and* (d): By the moment ratio lemma, the volume moments of the $(n-1)$-cone and $(n-1)$-cylinder about any axis through the vertex of the cone have ratio $2/n$. By the moment-volume principle (13.8), (with n replaced by $n-1$), the punctured and unpunctured cylindroids in n-space generated by rotating the cone and cylinder have the same volume ratio $2/n$. Combining this with Theorem 13.3a and 13.3b we obtain Theorem 13.3c and 13.3d.

Actually, Theorems 13.3a and 13.3b are equivalent because both the n-cylindroid and the punctured n-cylindroid are circumsolids, so by Theorem 4.13 their volumes have the same ratio as their surface areas. The same remark applies to Theorems 13.3c and 13.3d.

Note that the 2-cylindroid and the punctured 2-cylindroid are the same object, which is identical to their insphere (a disk), consistent with $2/n = 1$.

13.8 FURTHER EXTENSIONS TO n-SPACE, AND APPLICATIONS

Areas of spherical zones and cross sections.

Part (a) of the next theorem extends the relation illustrated in Figure 13.18b, and part (b) extends that in Figure 13.2a. The extensions refer to cross-sections of an n-cylindroid cut by an $(n-1)$-dimensional hyperplane perpendicular to the axis of symmetry of the cylindroid. A zone of a sphere or a cylindroid is the part of its surface between two parallel cross-sections.

Theorem 13.4. (a) *The surface areas of the corresponding zones of an n-sphere and its circumscribing n-cylindroid are equal.*

(b) *Corresponding cross-sections of an n-sphere and its punctured circumscribing n-cylindroid have equal areas.*

Proof of (a). In Figure 13.21b, the zone on the hemisphere is in area equilibrium with the projected slice of a circular disk. When they are rotated through 4-space around the balancing axis, the hemispherical zone produces a zone of a 4-sphere, and the projected slice produces a corresponding zone of the circumscribing 4-cylindroid. By the balance-revolution principle, the two zones have equal areas. This proves (a) when $n = 4$, and the same type of argument works for general n.

Proof of (b). Archimedes' chord-by-chord balancing of a triangle and semicircular disk in Figure 13.1a can be extended to 3-space for a cone and hemisphere obtained by rotating the diagram in Figure 13.1a about a horizontal axis.

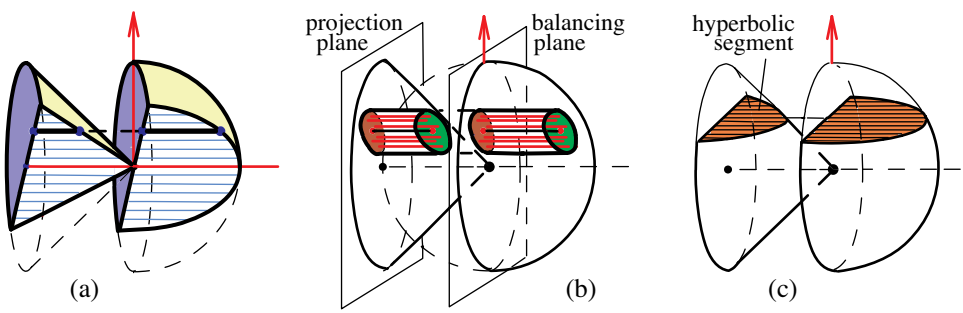


Figure 13.29: Extension of Archimedes' balancing in Figure 13.1a: (a) chords, (b) cables, (c) cross-sections of hemisphere and cone. The balance in (c) leads to equality of cross-sectional areas of a 4-sphere and its punctured circumscribing 4-cylindroid.

This balancing is indicated in Figure 13.29a, which shows a cross-section of the cone and hemisphere through the axis of symmetry. Puncture both the cone and hemisphere with an arbitrary collection of chords, as depicted by the examples in Figure 13.29b and 13.29c. Then the punctured parts will be in balance with respect to the central balancing plane. To see why, imagine two portions of a cable built from wires (the chords of the triangle and semicircle). Each pair of

wires is in balance, therefore their union (the two portions of the cable) are also in balance. In particular, each horizontal cross-sectional area of the cone will be in area balance with the corresponding cross-sectional area of the hemisphere, as illustrated in Figure 13.29c.

Rotate Figure 13.29c through 4-space, around a balancing axis perpendicular to the axis of symmetry and perpendicular to the 4th coordinate axis. The rotated cone sweeps out a punctured 4-cylindroid, and the hemisphere sweeps out a 4-sphere. Each cross-section of the 3-cone will sweep out a cross-section of the punctured 4-cylindroid, and the cross-section of the 3-hemisphere will sweep out the corresponding cross-section of the 4-sphere. Because the two cross-sectional areas are in balance before revolution, after revolution the swept cross-sectional hyperareas are equal according to the balance-revolution principle. This proves (b) when n=4, and the same type of argument works for general n.

Recursion formulas for volume and surface area of n-spheres.

Let V_n and S_n denote the volume and surface area, respectively, of an n-sphere of radius r. The following formulas are consequences of Theorem 13.3a and the moment-volume principle:

$$V_{n+1} = \frac{2\pi r^2}{n+1} V_{n-1}, \tag{13.10}$$

$$S_{n+2} = \frac{2\pi r^2}{n} S_n = 2\pi r V_n, \tag{13.11}$$

$$\frac{n+1}{n} \cdot \frac{V_{n+1}}{V_n} = \frac{V_{n-1}}{V_{n-2}}, \tag{13.12}$$

$$\frac{S_{n+2}}{S_{n+1}} = \frac{V_n}{V_{n-1}}. \tag{13.13}$$

Initial values are $V_1 = 2r, V_2 = \pi r^2, S_1 = 2, S_2 = 2\pi r$.

Proof. By Theorem 13.3a, V_n is equal to the volume of the punctured n-cylindroid of radius r, which, by the moment-volume principle, is $2\pi M_{n-1}(\text{cone})$, where $M_{n-1}(\text{cone})$ is the volume moment of the $(n-1)$-cone of altitude r that sweeps out the punctured n-cylindroid. Therefore, using (13.9) with $h = r$ we find

$$V_n = 2\pi M_{n-1}(\text{cone}) = \frac{2\pi r^2}{n} V_{n-2}.$$

Now replace n by $n+1$ to get (13.10) which, in turn, implies (13.12). The circumsolid property $V_n = (r/n)S_n$ in (4.16) yields (13.11). This implies $S_{n+2}/V_n = 2\pi r$, which is independent of n. Hence $S_{n+2}/S_{n+1} = V_n/V_{n-1}$, which is (13.13).

Using the notation $v_n = V_n/2$ and $s_n = S_n/2$ we see that the recursions (13.10) through (13.13) also hold for n-hemispheres, whose centroids will be determined in Section 13.9.

Volume and lateral surface area of n-dimensional cylindrical wedge.

The moment-wedge volume principle introduced in Section 13.4 for a cylindrical wedge can be extended to higher-dimensional space:

Moment-wedge volume principle. *The "volume" W_{n+1} of a general $(n+1)$-cylindrical wedge and the moment M_n of its n-spherical base are related by*

$$W_{n+1} = kM_n,$$

where k is a constant determined by the inclination of the truncating plane.

The moment-wedge volume principle has the corollary:

The "volumes" of two $(n+1)$-cylindrical wedges with the same k have the same ratio as the moments of their bases. In particular, the "volumes" are equal if the moments are equal.

We apply the corollary to the balanced n-hemisphere and n-cone shown in the top row of Figure 13.27. Instead of rotating these objects we use them as bases of truncated $(n+1)$-cylinders with the same height h. Their "volumes" are equal and can be calculated explicitly as follows. The truncated cylinder whose base is the n-cone is an $(n+1)$-pyramid whose base can be regarded as a "rectangle" of height h whose "area" is $V_{n-1}h$, where V_{n-1} is the "volume" of the $(n-1)$-sphere of radius r. The "volume" of this pyramid is $W_{n+1} = V_{n-1}h\cdot r/(n+1) = V_{n+1}h/(2\pi r)$, according to recursion (13.10). Replacing $n+1$ by n we get the theorem:

Theorem 13.5. *An n-cylindrical wedge of radius r and height h has volume W_n and lateral surface area A_n given by*

$$W_n = \frac{h}{2\pi r}V_n, \quad A_n = \frac{h}{2\pi r}S_n, \ n = 2, 3, \ldots,$$

where S_n is the surface area of an n-sphere of radius r.

When $n = 2$, the wedge is a triangle of area $W_2 = hr/2$ and altitude $A_2 = h$. When $n = 3$, $W_3 = (2/3)r^2h$, equivalent to the result obtained by Archimedes.

13.9 FORMULAS FOR CENTROIDS

This section applies our balancing relations to obtain formulas for locating centroids of various objects of interest.

1. Centroid of a regular circumgonal arc. A regular circumgon has an axis of symmetry through the center of its incircle, so the centroid also lies on this axis, as in Figure 13.30a. To find the distance c of the centroid from the incenter, rotate the circumgon so its symmetry axis is perpendicular to the balancing axis as in Figure 13.30b. Balancing means that $cL = rH$, where L is the length of the circumgonal arc and H is the length of its projection, hence

$$c = r\frac{H}{L}. \tag{13.14}$$

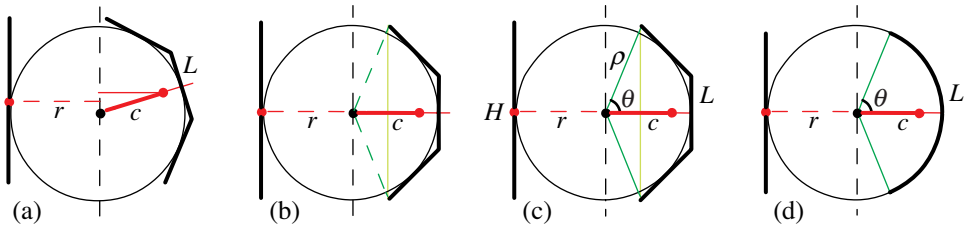


Figure 13.30: Rotating a regular circumgon in (a) to obtain a symmetric diagram in (b). (c) Relation between the centroidal distance c and the central angle subtended by the circumgon. (d) Limiting case when the circumgon becomes a circular arc.

From Figure 13.30c we see that if the regular circumgon consists of n edges and is subtended by a central angle 2θ, with ρ the distance to a vertex as indicated, then $H = 2\rho \sin\theta$. A simple exercise shows that $L = 2\rho n \sin(\theta/n)$, and (13.14) gives us

$$c = r\frac{\sin\theta}{n\sin\frac{\theta}{n}}. \tag{13.15}$$

The centroidal formula (13.14) for circumgonal arcs also holds for the limiting case when the circumgon is a circular arc of radius r with central angle 2θ, as in Figure 13.30d. In this case the projection has length $H = 2r\sin\theta$ and the circular arc has length $L = 2r\theta$, so the centroidal distance of the arc from the axis, which we denote by $c(\theta)$, is obtained from (13.14) as

$$c(\theta) = r\frac{\sin\theta}{\theta}. \tag{13.16}$$

This also follows from (13.15) when $n \to \infty$, and it was derived differently in Section 12.4, Example 8.

The next example uses (13.2) to determine area centroids.

2. Area centroid of a circumgonal region with inradius r. When the projection region is a triangle as in Figure 13.6c, we have $P = Hr/2$, $c_P = 2r/3$, and (13.2) yields

$$c_A = \frac{r^2}{3}\frac{H}{A} = \frac{2}{3}r\frac{H}{L}, \tag{13.17}$$

because $A = rL/2$. When the circumgon with central angle 2θ is rotated so its axis of symmetry is horizontal as in Figure 13.30c, the foregoing becomes

$$c_A = \frac{2}{3}r\frac{\sin\theta}{n\sin\frac{\theta}{n}}. \tag{13.18}$$

In the limiting case when the circumgonal region is a circular sector of radius r and central angle 2θ, the area centroid lies on the bisector of the sector at a distance $C(\theta)$ from the center, where

$$C(\theta) = \frac{2}{3}r\frac{\sin\theta}{\theta}. \tag{13.19}$$

This follows from (13.17) using $H = 2r\sin\theta$ and $L = r\theta$, or by letting $n \to \infty$ in (13.18). For a semicircular disk, $\theta = \pi/2$ and (13.19) gives $C(\pi/2) = 4r/(3\pi)$. This is the result Archimedes was seeking by balancing the triangle and semicircular disk in Figure 13.1b.

Next we use the double equilibrium in Figure 13.21a to treat the spherical wedge.

3. Area and volume centroids of a spherical wedge. There are two centroidal distances associated with the wedge in Figure 13.21a: the centroidal distance $C_A(\theta)$ for the spherical surface area A of the wedge, and $C_V(\theta)$ for the volume V of the wedge, where θ is the angle between the cutting plane and the equatorial plane. By symmetry, each of the centroids lies on the vertical plane bisecting the wedge.

Area centroid of a spherical wedge. Area equilibrium means that

$$C_A(\theta) \cdot A = r \cdot E, \tag{13.20}$$

where E is the area of the ellipse. But we have $A = 4\pi r^2(\theta/\pi) = 4\theta r^2$ and $E = \pi r H/2 = \pi r^2 \sin\theta$. Using these in (13.20) and solving for $C_A(\theta)$ we obtain

$$C_A(\theta) = \frac{\pi}{4} r \frac{\sin\theta}{\theta}. \tag{13.21}$$

Volume centroid of a spherical wedge. To find the distance of the volume centroid from the balancing plane we use the volume equilibrium relation

$$C_V(\theta) \cdot V = \frac{3}{4} r \cdot V_E, \tag{13.22}$$

where V_E is the volume of the elliptical cone. In this case we have $V = 4\theta r^3/3$ and $V_E = Er/3 = \pi r^2 H/6 = (\pi/3)r^3 \sin\theta$, and when these are used in (13.22) we find

$$C_V(\theta) = \frac{3\pi}{16} r \frac{\sin\theta}{\theta}. \tag{13.23}$$

4. Volume centroid of a spherical sector and of a spherical segment. In Figure 13.20a, volume equilibrium about the balancing plane states that

$$C(\theta)V_s = \frac{3}{4} r V_c, \tag{13.24}$$

where $C(\theta)$ is the centroidal distance from the balancing plane, V_s is the volume of the spherical sector, and V_c is the volume of the cone. By (13.4) we have $V_s = 2\pi r^2 h/3$, where $h = r - r\cos\theta$ is the height of the spherical cap of the sector. The volume of the cone is $V_c = \pi(r\sin\theta)^2/3$. Using these in (13.24) and solving for $C(\theta)$ we find

$$C(\theta) = \frac{3}{8} r(1 + \cos\theta). \tag{13.25}$$

When the conical part is removed from the spherical sector in Figure 13.20a, the solid that remains is called a spherical segment. Archimedes in [47; *Method*, Prop.

9, p. 35] determines the volume centroid of a spherical segment. His description for the distance c of the centroid from the center can be stated alternatively as

$$c = \frac{3}{4}\frac{(r+h)^2}{2r+h}, \tag{13.26}$$

where $h = r\cos\theta$ is the altitude of the conical part of the sector of radius r.

The proof follows from the balancing relation involving volumes:

$$V_{\text{cone}} \cdot c_{\text{cone}} + V_{\text{segm}} \cdot c_{\text{segm}} = V_{\text{sect}} \cdot c_{\text{sect}}. \tag{13.27}$$

We know that $V_{\text{cone}} = \pi h(r^2 - h^2)/3$, $c_{\text{cone}} = 3h/4$, $V_{\text{sect}} = 2\pi r^2(r-h)/3$, and from (13.25), $c_{\text{sect}} = 3(r+h)/8$. Also, $V_{\text{segm}} = V_{\text{sect}} - V_{\text{cone}} = \pi(r-h)^2(2r+h)/3$. When these are used in (13.27) we obtain (13.26). In the special case when $h = 0$, (13.26) yields $c = 3r/8$ for the volume centroid of a hemisphere, which also follows from (13.23) with $\theta = \pi/2$.

5. Centroids of an n-hemisphere. Let v_n, s_n denote, respectively, volume and surface area of an *n-hemisphere* of radius r, and let $c(v_n)$, $c(s_n)$ denote their respective centroidal distances from the center. Then we have the following remarkable theorem:

Theorem 13.6. *The following recursions hold:*

$$c(v_n) = \frac{n}{n+1}c(v_{n-2}) \quad \text{and} \quad c(s_{n+2}) = \frac{n}{n+1}c(s_n), \tag{13.28}$$

with initial values

$$c(v_1) = r/2,\ c(v_2) = 4r/(3\pi),\ c(s_1) = r,\ c(s_2) = 2r/\pi.$$

Proof. When an n-hemisphere of volume v_n and surface area s_n is rotated about a diameter, it generates an $(n+1)$-sphere of volume $2v_{n+1}$ and surface area $2s_{n+1}$. By the moment-volume principle (13.8) we have $2v_{n+1} = 2\pi v_n c(v_n)$ and $2s_{n+1} = 2\pi s_n c(s_n)$, from which we find

$$c(v_n) = \frac{v_{n+1}}{\pi v_n} \quad \text{and} \quad c(s_n) = \frac{s_{n+1}}{\pi s_n} = \frac{v_{n-1}}{\pi v_{n-2}},$$

the last relation coming from (13.13). Using (13.10) in these and invoking (13.12) we find

$$\frac{c(v_n)}{c(v_{n-2})} = \frac{c(s_{n+2})}{c(s_n)} = \frac{v_{n+1}}{v_{n-1}} \cdot \frac{v_{n-2}}{v_n} = \frac{n}{n+1}.$$

This proves (13.28). Joint recursion (13.13) together with (13.28) yields:

Corollary of Theorem 13.6.

$$c(s_{n+2}) = c(v_n). \tag{13.29}$$

This surprising and remarkable result equates the centroid of an $(n+2)$-dimensional surface with that of an n-dimensional volume. When $n = 1$, it tells us that the area

centroid of a 3-dimensional hemisphere is at a distance $r/2$ from its center because a 1-dimensional hemisphere is a line segment with its centroid at its midpoint.

Repeated use of (13.29) leads to the explicit formulas for $n \geq 1$:

$$c(s_{n+2}) = c(v_n) = \frac{n!!}{(n+1)!!} p_n, \tag{13.30}$$

where $p_n = c(s_1) = r$ for odd n, $p_n = c(s_2) = 2r/\pi$ for even n, and $n!!$ is the double factorial symbol. Both $c(v_n)$ and $c(s_n)$ are rational multiples of r for odd n, and of r/π for even n.

It is easy to see that the following unexpected reciprocal recursions also hold:

$$c(v_{n+1}) = \frac{2r^2}{\pi(n+2)} \frac{1}{c(v_n)}, \quad c(s_{n+1}) = \frac{2r^2}{\pi n} \frac{1}{c(s_n)}. \tag{13.31}$$

Unlike those in (13.28), they relate centroidal distances in consecutive dimensions.

6. Centroid of an n-spherical zone. It is well known that the centroid of the surface of a slice of a sphere is midway between the two parallel cutting planes. (See Corollary 5.6.) This is also seen by referring to Figure 13.31, noting that from

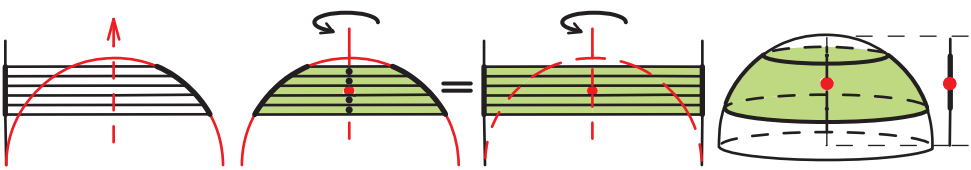


Figure 13.31: Centroid of a spherical zone is midway between the cutting planes.

(13.3) the area of a spherical zone grows linearly with its height H. When the zone is a spherical cap as in Figure 13.19b, its area centroid is at the midpoint of the height of the cap. In particular, when the cap is a full hemisphere of radius r, the area centroid is at distance $r/2$ from the center. This also follows by taking $n = 1$ in (13.29). Now we will show that (13.29) holds not only for n-hemispheres but for any n-spherical zone. This recursion is included as part (c) of the following theorem.

Theorem 13.7. (a) *The zones in* Theorem 13.4a*, which have equal areas, also have equal centroidal altitudes (above the common equatorial hyperplane).*

(b) *The volume of a slice of a solid n-sphere and the surface area of the corresponding $(n+2)$-cylindroidal zone, which it generates, have equal centroidal altitudes.*

(c) *The surface area of an $(n+2)$-spherical zone and the volume of the corresponding slice of an n-sphere of the same radius have equal centroidal altitudes.*

Proof. The proof of (a) follows by analogy with Figure 13.31, with the cylinder replaced by a cylindroid. We regard each zone as being made up of thin elemental zones. Corresponding elements have equal areas and equal centroidal altitudes, so the same is true for the full zone. Part (b) follows from the fact that when an object is rotated around an axis, each of its points stays at the same altitude above a fixed

hyperplane perpendicular to the rotation axis. Part (c) follows from (a) and (b) by eliminating the common cylindroid.

Part (c) can be expressed as a recursion formula. Let $c(s_n^{slice})$ denote the centroidal altitude of the surface area of a zone of an n-sphere between two parallel planes, and let $c(v_n^{slice})$ denote the centroidal altitude of the volume of the corresponding slice of the solid n-sphere of the same radius between the same parallel planes. Part (c) states that

$$c(s_{n+2}^{slice}) = c(v_n^{slice}). \tag{13.32}$$

When the slice is a full hemisphere we obtain recursion (13.29). Again, we note the remarkable nature of this recursion. It equates the centroid of an $(n + 2)$-dimensional surface with that of an n-dimensional volume. The reason for this phenomenon is that the cylindroid that occurred in parts (a) and (b) of Theorem 13.7 was eliminated to obtain part (c).

7. Centroid of a Lambert-type projection on an n-cylindroid. Theorem 13.7a can be extended so that it applies not only to zones but also to any region of the sphere and its special projection onto the lateral surface of the circumscribing cylindroid. In 3-space this special projection is known as Lambert's mapping (discussed in Section 5.12). It projects the surface of a sphere onto a lateral circumscribing cylinder by rays through the polar axis that are parallel to the equatorial plane. In Section 5.12 we showed that Lambert's mapping preserves areas. Moreover, the regions of equal area on the sphere and cylinder have the same centroidal altitude above the equatorial plane. To see this, refer to Figure 13.18a, which shows an elemental circular arc Δs and its balanced projection Δp. If the projection line is moved to the opposite side of the balancing axis, the elements Δp and Δs will still have equal moments about the balancing axis. Now rotate both through a small angle around the balancing axis. The rotated arc produces a surface element on the sphere and its rotated projection produces a surface element on the cylinder that is a Lambert projection of the spherical element. According to the balance-revolution principle, the two elements have equal areas. They also have the same centroidal altitude above the equatorial plane. Because this holds for all such surface elements, these two results are also true for any region on the sphere and its Lambert projection on the cylinder. This same type of proof also works in n-space, with the cylinder replaced by a circumscribing n-cylindroid. The polar axis is replaced by the balance-revolution axis that produces the cylindroid as in Figure 13.27. Note that this gives another proof of Theorem 13.7a because a cylindroidal zone is a Lambert-type projection of its corresponding spherical zone.

13.10 ON THE SPHERE AND ITS CIRCUMSOLIDS IN n-SPACE

Archimedes' relation on the volume and surface area of a cylinder and its insphere consists of two parts: (1) their volumes and surface areas have the same ratio, and (2) this ratio, cylinder to sphere, is 3/2. Theorem 4.13 tells us that part (1) is true not only for the cylinder and its insphere but for any two circumsolids having the same insphere. When one circumsolid is the insphere itself, Theorem 4.13 implies:

Theorem 13.8. *For an n-sphere and any of its n-circumsolids, their volumes and outer surface areas have the same ratio.*

Theorem 13.8 extends part (1). This rest of this section extends part (2) by investigating the common ratio, which we denote by $\rho(n)$. Because of Theorem 13.8 we treat the ratio for volumes:

$$\rho(n) = \frac{V_n(\text{circumsolid})}{V_n(r)}, \tag{13.33}$$

where $V_n(r)$ is the volume of the insphere of radius r of a given n-circumsolid. Our primary interest is the behavior of $\rho(n)$ as a function of the dimension n.

Using (13.8) with n replaced by $n-1$, together with (13.9) with $h = r$, we find that the volume of an n-cylindroid bears a simple relation to the volume $V_{n-2}(r)$ of a sphere having 2 lower dimensions:

$$V_n(\text{cylindroid}) = \pi r^2 V_{n-2}(r). \tag{13.34}$$

This relation, together with recursion (13.10), gives us the elementary ratio

$$\rho(n) = \frac{\pi r^2 V_{n-2}(r)}{V_n(r)} = \frac{n}{2}, \tag{13.35}$$

which also emerges from Theorem 13.3. This $\rho(n)$ has such a simple form because it depends on the ratio of volumes of spheres whose dimensions differ by 2, and recursion (13.10) simplifies the ratio to yield (13.35).

For the next three circumsolids, $\rho(n)$ depends on the ratio of volumes of spheres of *consecutive* dimensions, for which there is no simple recursion akin to (13.10). Before discussing these circumsolids, we investigate general properties of the ratio $V_{n-1}(r)/V_n(r)$.

Properties of the ratio $V_{n-1}(r)/V_n(r)$.

It is known (see [2; p. 411]) that the volume of an n-sphere of radius r can be expressed explicitly in terms of the gamma function:

$$V_n(r) = \frac{\pi^{n/2}}{\Gamma(\frac{n+2}{2})} r^n. \tag{13.36}$$

Thus, $V_n(r)$ is proportional to r^n, but the proportionality factor is a nonelementary function of n. The ratio of volumes of spheres is essentially the ratio of two gamma functions. In particular,

$$\frac{rV_{n-1}(r)}{V_n(r)} = \frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{n+2}{2})}{\Gamma(\frac{n+1}{2})}. \tag{13.37}$$

There is a corresponding formula for the ratio $r^2 V_{n-2}(r)/V_n(r)$, and in this case recursion (13.10) tells us that the ratio of gamma functions simplifies and leads to (13.35). No such simplification occurs for the ratio in (13.37). However, we have

an interesting geometric interpretation of the ratio that makes no use of gamma functions. From (13.30) we find

$$\frac{V_{n-1}(r)}{V_n(r)} = \frac{1}{\pi c(v_{n-1})}, \tag{13.38}$$

where $c(v_{n-1})$ is the centroidal distance of an $(n-1)$-hemisphere from its center.

No simple formula reveals how $c(v_{n-1})$ behaves as a function of n, but it is easy to determine its asymptotic behavior for large n. From recursion (13.31) we find $c(v_{n-2})c(v_{n-1}) = 2r^2/(\pi n)$, which leads to the asymptotic relation for large n:

$$c(v_{n-1}) \sim r\sqrt{\frac{2}{\pi n}}. \tag{13.39}$$

Using this in (13.38) we obtain

$$\frac{rV_{n-1}(r)}{V_n(r)} \sim \sqrt{\frac{n}{2\pi}} \tag{13.40}$$

for large n. The asymptotic formula can also be deduced in a nonelementary way by applying Stirling's formula for the gamma function ratio in (13.37).

Circumsolids of inradius r for which $\rho(n)$ depends on the ratio $V_{n-1}(r)/V_n(r)$.

1. n-cylinder of radius r and altitude $2r$. The volume of the n-cylinder is $2rV_{n-1}(r)$, so (13.33) gives us

$$\rho(n) = \frac{2rV_{n-1}(r)}{V_n(r)}. \tag{13.41}$$

When $n = 3$ this yields $\rho(3) = 3/2$, the classical result of Archimedes.

Using (13.40) in (13.41) we obtain, for an n-cylinder, the asymptotic value

$$\rho(n) \sim \sqrt{\frac{2n}{\pi}} \tag{13.42}$$

for large n. By contrast, an n-cylindroid has the exact value $\rho(n) = n/2$ for all n.

2. n-double cone of vertex angle 2α. This double cone is the union of two congruent cones, each having base of radius $R = r/\cos\alpha$, altitude $H = r/\sin\alpha$, and volume

$$V_{n-1}(R)\frac{H}{n} = (\frac{R}{r})^{n-1}V_{n-1}(r)\frac{r}{n\sin\alpha} = \frac{r}{n\sin\alpha}(\frac{1}{\cos\alpha})^{n-1}V_{n-1}(r).$$

In this case (13.33) gives

$$\rho(n) = \frac{1}{n\sin\alpha}(\frac{1}{\cos\alpha})^{n-1}\frac{2rV_{n-1}(r)}{V_n(r)}. \tag{13.43}$$

For a 3-dimensional circumsolid double cone with $\alpha = \pi/4$, (13.43) gives $\rho(3) = \sqrt{2}$. This is equivalent to Archimedes' discovery that such a cone *inscribed* in a sphere, as in Figure 13.36, has volume half that of its circumscribing sphere.

3. n-cone of vertex angle 2α. The altitude of this cone is $H = r + r/\sin\alpha = r(1+\sin\alpha)/\sin\alpha$ and the radius of its base is $R = H\tan\alpha = r(1+\sin\alpha)/\cos\alpha$, hence its volume is

$$V_{n-1}(R)\frac{H}{n} = (\frac{R}{r})^{n-1}V_{n-1}(r)\frac{r}{n}\frac{1+\sin\alpha}{\sin\alpha} = \frac{r}{n\tan\alpha}(\frac{1+\sin\alpha}{\cos\alpha})^n V_{n-1}(r). \quad (13.44)$$

For this example, (13.33) yields

$$\rho(n) = \frac{(1+\sin\alpha)^n}{n\tan\alpha\cos^n\alpha}\frac{rV_{n-1}(r)}{V_n(r)}. \quad (13.45)$$

4. Truncated n-cone of vertex angle 2α. Truncate the n-cone of altitude H in the foregoing example by removing a smaller cone of altitude h with base of radius R_1 tangent to the insphere. The volume of the smaller cone is $V_{n-1}(R_1)(h/n) = (R_1/r)^{n-1}V_{n-1}(r)(h/n)$. Using $h = H - 2r = r(1-\sin\alpha)/\sin\alpha$ and $R_1 = h\tan\alpha = r(1-\sin\alpha)/\cos\alpha$, we find that the smaller cone has volume

$$\frac{r}{n\tan\alpha}(\frac{1-\sin\alpha}{\cos\alpha})^n V_{n-1}(r). \quad (13.46)$$

The volume in (13.44) minus that in (13.46) is the volume of the truncated cone n-circumsolid, for which (13.33) yields

$$\rho(n) = \frac{(1+\sin\alpha)^n - (1-\sin\alpha)^n}{n\tan\alpha\cos^n\alpha}\frac{rV_{n-1}(r)}{V_n(r)}. \quad (13.47)$$

When $\alpha = \pi/6$ we find the values $\rho(2) = 8/(\pi\sqrt{3})$ and $\rho(3) = 13/6$.

We turn next to examples of n-circumsolids whose volumes, like that of the n-cylindroid, are simply related to $V_{n-2}(r)$.

Circumsolids for which $\rho(n)$ depends on the elementary ratio $V_{n-2}(r)/V_n(r)$.

5. n-double conoid circumscribing n-sphere. Figure 13.32, which is analogous to Figure 13.28, helps visualize the construction of such an n-dimensional double conoid. Start with an $(n-1)$-cone of radius r and altitude r inscribed in an $(n-1)$-hemisphere of radius r. It is also inscribed in an $(n-1)$-cylinder of radius r and altitude r. The volume of the cylinder is $r \cdot V_{n-2}(r)$ and that of the cone is $V_{n-1}(\text{cone}) = r \cdot V_{n-2}(r)/(n-1)$. The centroid of the cone is at a distance r/n from the center. When the cone is tumbled in n-space about an axis through the diameter of its base, it sweeps out a solid in n-space that we call an n-double inconoid because it is *inscribed* in an n-sphere. Its insphere has radius $r/\sqrt{2}$, and its volume is

$$V_n(\text{inconoid}) = 2\pi\frac{r}{n}V_{n-1}(\text{cone}) = \frac{2\pi r^2}{n(n-1)}V_{n-2}(r) = \frac{V_n(r)}{n-1}, \quad (13.48)$$

the last relation coming from (13.10). To obtain a solid circumscribing the n-sphere of radius r, expand the inconoid and its insphere radially by the factor $\sqrt{2}$. We call the resulting circumsolid an n-double conoid, whose volume is $V_n(\text{circumsolid}) =$

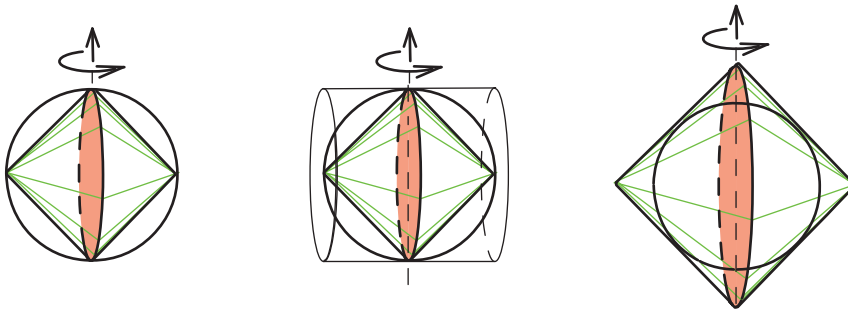


Figure 13.32: Construction of a double conoid in 4-space by analogy to Figure 13.28.

$2^{n/2}V_n(r)/(n-1)$. For this example,

$$\rho(n) = \frac{2^{n/2}}{n-1}$$

in (13.33), which gives $\rho(3) = \sqrt{2}$, in agreement with the result obtained above in Example 2 with $\alpha = \pi/4$.

6. *n*-hexaconoid circumscribing *n*-sphere. This circumsolid is of special interest because it leads to a ratio $\rho(n)$ that is rational for every $n \geq 3$. Figure 13.33 shows a special circumsolid of revolution formed by rotating a regular hexagon about a diameter of a sphere of radius r. Half of the circumsolid is composed of two parts, a circular cylinder of radius r and altitude $h = r/\sqrt{3}$, and an adjacent cone of radius r with the same altitude h. We start with this solid, which we call an n-hexacone and build an $(n+1)$-hexaconoid by analogy with the construction used for the cylindroid in the schematic representation in Figure 13.28. An $(n+1)$-hexaconoid

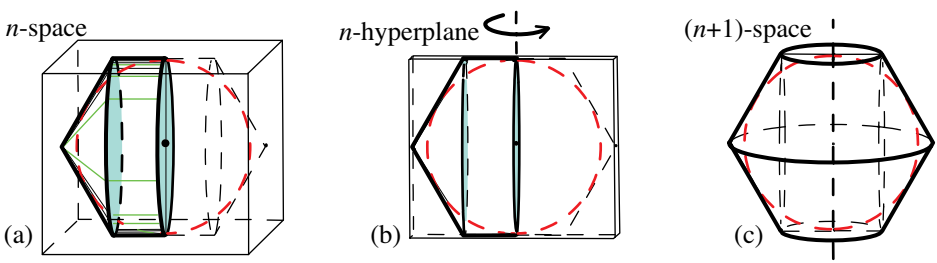


Figure 13.33: Constructing an $(n+1)$-hexaconoid. (a) Solid hexacone in n-space. (b) Schematic flattening of (a) onto a hyperplane, with a new axis erected perpendicular to it. (c) Rotation of (b) in $(n+1)$-space around a coordinate axis perpendicular to this new axis and to the axis of symmetry.

is constructed from an n-hexacone in an analogous fashion. The n-hexacone consists of two parts, an n-cylinder plus an adjacent n-cone of the same radius r and altitude $h = r/\sqrt{3}$. Their volumes are related by $V_n(\text{hexacone}) = V_n(\text{cylinder}) + V_n(\text{cone})$.

Tumbling the n-hexacone about an axis in $(n+1)$-space produces an $(n+1)$-hexaconoid whose volume, by the moment-volume principle in (13.8), is given by

$$V_{n+1}(\text{hexaconoid}) = 2\pi M_n(\text{hexacone}),$$

where $M_n(\text{hexacone})$ is the moment of the n-hexacone about that axis. Moments are additive, so

$$M_n(\text{hexacone}) = M_n(\text{cylinder}) + M_n(\text{cone}). \tag{13.49}$$

To calculate the moments on the right of (13.49) we note that $V_n(\text{cylinder}) = hV_{n-1}(r)$, where $V_{n-1}(r)$ is the volume of an $(n-1)$-sphere of the same radius; its moment is $h/2$ times this volume, so $M_n(\text{cylinder}) = (h^2/2)V_{n-1}(r)$. Similarly, $V_n(\text{cone}) = (h/n)V_{n-1}(r)$ and $M_n(\text{cone}) = (h^2/n)V_{n-1}(r)(n+2)/(n+1)$, so (13.49) gives $M_n(\text{hexacone}) = \frac{r^2}{6}(1 + \frac{2}{n}\frac{n+2}{n+1})V_{n-1}(r)$. From (13.38) and (13.10) we find

$$\rho(n+1) = \frac{V_{n+1}(\text{hexaconoid})}{V_{n+1}(r)} = \frac{2\pi M_n(\text{hexacone})}{V_{n+1}(r)} = \frac{n^2+3n+4}{6n}. \tag{13.50}$$

When $n = 2$ this yields $\rho(3) = 7/6$, and when $n = 3$ it gives $\rho(4) = 11/9$.

Figure 13.34 shows how to generate another type of hexaconoid. Start with the same n-hexacone as in Figure 13.33a but orient it so that it stands on a vertex as in Figure 13.34a. When this is tumbled in higher-dimensional space it generates the $(n+1)$-hexaconoid shown in Figure 13.34b. A calculation similar to that producing (13.50) yields

$$\rho(n+1) = \frac{2^{n+1} - n - 2}{n \cdot 3^{(n-1)/2}}.$$

When $n = 2$ this gives $\rho(3) = 2/\sqrt{3}$, and when $n = 3$ it gives $\rho(4) = 11/9$, which agrees with (13.50) (because the two hexaconoids are the same object in 4-space).

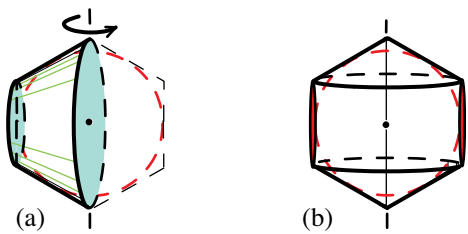


Figure 13.34: Another way to generate another type of n-hexaconoid.

A special family of circumsolids.

Recall the family of punctured 3-dimensional circumsolids in Figure 13.14. They are right circumsolids, obtained by rotating half a regular $2n$-gonal region of the type shown in Figure 13.9. Each has volume $4\pi r^3/3$, which is that of their common insphere of radius r. The corresponding unpunctured solids have a larger volume, which is $4\pi r^3/3$ plus the volumes of the two cones that puncture the solids. Direct

calculation of the volume of each puncturing cone generated by a regular $2n$-gon reveals that the total volume of a typical unpunctured solid is given by

$$\text{Volume} = \frac{4}{3}\pi r^3 + \frac{2}{3}\pi r^3 \tan^2(\frac{\pi}{2n}). \tag{13.51}$$

When $n \to \infty$ the second term in (13.51) tends to 0 and we obtain the volume of the insphere as the limiting value, as expected. Incidentally, the total surface area of each unpunctured solid is $3/r$ times its volume.

For this family, the ratio of the volume of the circumgonal solid to that of its insphere is given by

$$\frac{\text{Volume}}{\frac{4}{3}\pi r^3} = 1 + \frac{1}{2}\tan^2(\frac{\pi}{2n}).$$

This ratio is also that of the corresponding total surface areas. When $n = 2$, the circumsolid is a cylinder and the ratio is $3/2$, the same result obtained by Archimedes. When $n = 3$ the circumsolid is obtained by rotating a regular hexagon, and the ratio is $7/6$, the same result obtained by taking $n = 2$ in (13.50). For large n the ratio is asymptotic to $1 + \pi^2/(8n^2)$.

We can use the circumsolids in this family to construct a family of circumsolids in higher-dimensional space by the same method we used to construct the cylindroid. Figure 13.35 shows a rearrangement of the unpunctured circumsolids in Figure 13.14. Tumbling the lower half of the cylinder generates a 4-cylindroid. Tumbling the lower half of the other circumsolids generates a family of 4-circumsolids. Repeated rotation of these solids around the vertical axis of symmetry produces higher-dimensional solids, such as cylinder, hexacone, etc. Tumbling these about an axis through the equatorial diameter produces a cylindroid, hexaconoid, etc. Each of these circumsolids gives a value of $\rho(n)$ proportional to the ratio $V_{n-2}(r)/V_n(r)$. The constant of proportionality will depend on the parameters defining the particular circumsolid of the family.

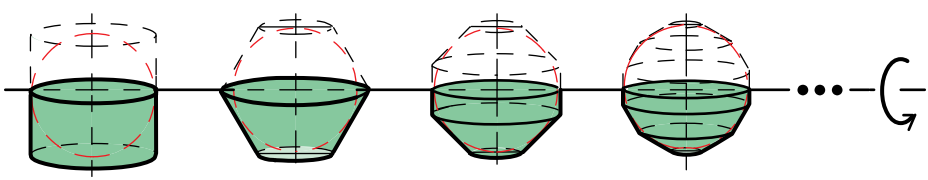


Figure 13.35: Constructing circumsolids with $\rho(n)$ proportional to $V_{n-2}(r)/V_n(r)$.

Note that the double conoid in Figure 13.32 and the hexaconoid of the second type in Figure 13.34 can also be obtained from the first two regular polygons in Figure 13.35 (the square and hexagon) by orienting them so they stand on a vertex rather than on the base. The same can be done with the other polygons in Figure 13.35. Thus, there are infinitely many circumsolids for which $\rho(n)$ is proportional to $V_{n-2}(r)/V_n(r)$. But the cylindroid stands out because its ratio $\rho(n) = n/2$ is so simple and yields the Archimedes ratio $3/2$ when $n = 3$.

Miscellaneous observations.

We conclude this section with miscellaneous remarks relating some of the results in this chapter to results found by Archimedes.

1. *n*-sphere and its *n*-double inconoid. In his discovery of the 3/2 ratio for both the volumes and surface areas of a cylinder and its insphere, Archimedes never compared the sphere directly with its circumscribing cylinder. In [47; *Method*, Prop. 2, p. 18] he used mechanical balancing to discover that the volume of a sphere is four times that of an inscribed cone whose base is a great circle of the sphere and whose altitude is equal to its radius. In [47; *On the Sphere and Cylinder, Book I,* Prop. 34, p. 41] he proved this result using the method of exhaustion. That cone and its mirror image form the double cone in Figure 13.36a. Archimedes' Proposition 34 states that the volume of the sphere is exactly twice that of the double inscribed cone. He knew that the volume of the double cone, in turn, is one-third that of the circumscribing cylinder in Figure 13.36a, and deduced, as a corollary of Proposition 34, that the volume of the cylinder is 3/2 that of the sphere. Equation (13.48) provides a higher-dimensional generalization of Proposition 34:

Theorem 13.9. *The volume of an n-sphere is* $(n-1)$ *times that of the n-double inconoid.*

In Proposition 33, p. 39, Archimedes proved that the surface area of a sphere is four times the area of the greatest circle in it. He combined this with Proposition 13, p. 16 (regarding the lateral surface area of a cylinder), and concluded that the total surface area of the circumscribing cylinder (including the bases) is 3/2 times that of the sphere, the same ratio as their volumes. He was so excited by this discovery that he wanted a cylinder and inscribed sphere engraved on his tombstone.

2. *n*-cube and its insphere. Archimedes must have realized that the volume ratio of a cube circumscribing a sphere has the same value as the ratio of their surface areas, which, in modern terminology, is $6/\pi$. This ratio is not simple like 3/2 and probably would not have generated much excitement for Archimedes.

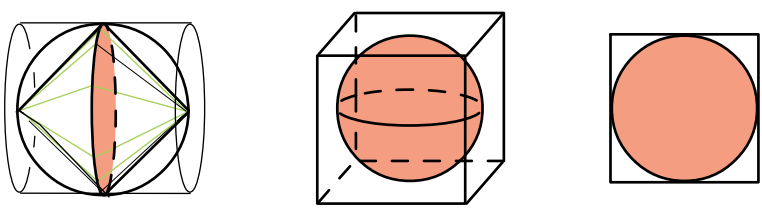


Figure 13.36: (a) Double cone considered by Archimedes. (b) Cube circumscribing 3-sphere. (c) Square circumscribing 2-sphere, a circular disk.

An n-cube of edge-length $2r$ is a circumsolid with n-insphere of radius r. The n-cube has volume $(2r)^n$, so the ratio in (13.33) is $\rho(n) = (2r)^n / V_n(r)$. The elementary formulas $V_3(r) = 4\pi r^3/3$ and $V_2(r) = \pi r^2$ yield $\rho(3) = 6/\pi$ and $\rho(2) = 4/\pi$,

illustrated in Figures 13.36b and c. For general n, $V_n(r)$ is given by (13.36), and

$$\rho(n) = \frac{2^n \Gamma(\frac{n+2}{n})}{\pi^{n/2}}.$$

We remind the reader that $\rho(n)$ also represents the ratio of the total surface area of the n-cube to the total surface area of its n-insphere.

3. Square and its incircle. Many ancient civilizations observed through measurement that the ratio of the circumference of a circle to its diameter is the same for all circles. The Greeks were the first to explain that this is a simple consequence of similarity. A circular disk of radius r is similar to a unit circular disk (of radius 1), the scaling factor being r. If a unit disk has circumference C and area A, then a disk of radius r has circumference Cr and area Ar^2 (because lengths are multiplied by the scaling factor, and areas by the square of the scaling factor). Thus, the two numbers C and A represent fundamental constants that tell us how to determine the circumference and area of a circular disk in terms of its radius. The ratio of circumference to diameter is $(Cr)/(2r) = C/2$, a constant independent of the radius. Today this ratio is denoted by the symbol π. Thus, $C = 2\pi$ and a disk of radius r has circumference $Cr = 2\pi r$.

Archimedes' discovery in [47; *Measurement of Circle*, Prop. 1] is equivalent to the statement that the area of a circular disk is one half its circumference times its radius. For a unit disk this states that $A = C/2$, a landmark discovery relating these fundamental constants. Thus, the area of a circular disk of radius r is πr^2. This is equivalent to $\rho(2) = 4/\pi$ for a square and its incircle.

4. Rational ratios and Archimedes. Archimedes realized the importance of the ratio of circumference to diameter. In [47; *Measurement of Circle*, Prop. 3] he obtains various rational approximations to it by comparing the circumference of a circle with the perimeters of inscribed and circumscribed regular polygons. He starts with regular hexagons and keeps doubling the number of sides, until he reaches polygons with 96 sides and obtains Proposition 3, which is equivalent to the inequalities

$$3\tfrac{10}{71} < \pi < 3\tfrac{1}{7}.$$

We speculate that Archimedes realized that π, like $\sqrt{2}$, is not a rational number, although he had no proof. A proof of the irrationality of $\sqrt{2}$ appears in Euclid's *Elements*, but the irrationality of π was first established nearly 2000 years later by Johann Lambert in 1761.

It is clear that Archimedes was excited by geometric measurements leading to rational ratios: The 2/3 ratio of the volumes and surface areas of a sphere and cylinder, and the 2/3 ratio of the volume of a cylindrical wedge to its circumscribing prism. He was not aware that their lateral surface areas also have the same ratio.

5. Extension to volume of n-ellipsoid. If an n-sphere together with its circumscribing cylindroid is dilated by an arbitrary factor along each of its n coordinate axes, we obtain a general n-ellipsoid with its circumscribing ellipsoidal cylindroid. The dilations preserve volume relations but not surface area relations. Therefore,

parts (a) and (c) of Theorem 13.3 also hold for the ellipsoids, namely, equality of volumes with punctured cylindroids, and the $2/n$ ratio with the unpunctured cylindroid. When $n = 3$ they include results found by Archimedes for ellipsoids of revolution.

There is also an extension of Theorem 13.4b for equality of corresponding cross-sectional areas of an n-ellipsoid and its punctured circumscribing cylindroid. This implies equality of volumes of slices between corresponding parallel hyperplanes.

NOTES ON CHAPTER 13

Much of this chapter is contained in a paper entitled "New balancing principles applied to circumsolids of revolution, and to n-dimensional spheres, cylindroids and cylindrical wedges." The paper was accepted in 2012 for publication in the *American Mathematical Monthly.*

Chapter 14

SUMS OF SQUARES

These problems can be easily solved by the methods developed in this chapter. The reader may wish to try solving them before reading the chapter.

1. Given two distinct points, the locus of all points in a plane containing them such that the sum of distances from them is constant is known to be an ellipse, with the given points as foci.

(a) *What is the locus of points in the plane if the sum of squares of distances from the given points is constant?*

(b) *Determine the locus if we start with any number of distinct points in 3-space.*

2. The following identities involve sums of squares of consecutive integers and can be regarded as extensions of the Pythagorean triple in the first formula:

$$3^2 + 4^2 = 5^2$$
$$10^2 + 11^2 + 12^2 = 13^2 + 14^2$$
$$21^2 + 22^2 + 23^2 + 24^2 = 25^2 + 26^2 + 27^2.$$

(a) *Find an infinite family of identities involving squares of consecutive integers that include the foregoing as special cases.*

Each of the following identities involves sums of squares of consecutive integers with alternating signs:

$$-5^2 + 4^2 = -3^2$$
$$2^2 - 3^2 + 4^2 = -5^2 + 6^2$$
$$-11^2 + 10^2 - 9^2 + 8^2 = -7^2 + 6^2 - 5^2.$$

(b) *Find an infinite family of identities involving squares of consecutive integers with alternating signs that include the foregoing as special cases.*

CONTENTS

14

This chapter begins with a geometric problem concerning the locus of a point with the sum of squares of distances from n fixed points being constant. When generalized to higher-dimensional space, the problem leads to several remarkable formulas for sums of squares of distances in a finite-dimensional space. One of them is a discrete analog of the parallel-axis theorem and its higher-dimensional extension. Another is to intrinsic second moments of composite systems. After providing several applications to physics and geometry, the chapter changes emphasis in the last section, which gives a simple method for discovering families of striking identities involving sums of squares of consecutive integers and, more generally, of integers in arithmetic progression.

14.1 A LOCUS PROBLEM IN THE PLANE

A point that moves in a plane so that the sum of its distances from two fixed points in the plane is constant traces an ellipse with the two points as foci.

What is the locus if the point moves so that the sum of the squares of the two distances is constant?

An elementary calculation in coordinate geometry shows that the locus is a circle with center midway between the two points.

When we ask the same question for three or more fixed points in a plane we obtain the following surprising result.

Theorem 14.1. *Given n fixed points in a plane, a point moving in the plane so that the sum of the squares of the distances from the points is constant traces out a circle whose center is at the centroid of the fixed points.*

Proof. Start with n arbitrary points in a plane, and let O denote their centroid. Using O as the origin of a complex plane, denote the points by the n complex numbers $z_1, z_2, \ldots, z_n$, so that

$$\sum_{k=1}^{n} z_k = 0. \tag{14.1}$$

Now let z denote an arbitrary point in the plane, and consider the sum of the squares of the distances from z to the points:

$$\sum_{k=1}^{n} |z - z_k|^2.$$

The kth term of the sum is

$$(z - z_k)(\overline{z} - \overline{z}_k) = |z|^2 + |z_k|^2 - z\overline{z}_k - \overline{z}z_k.$$

Summing on k and using (14.1) we find

$$\sum_{k=1}^{n} |z - z_k|^2 = n|z|^2 + \sum_{k=1}^{n} |z_k|^2 = n|z|^2 + nD_n^2, \tag{14.2}$$

where

$$D_n^2 = \frac{1}{n} \sum_{k=1}^{n} |z_k|^2$$

is the average of the squares of the distances of the points z_1, z_2, ..., z_n from their centroid. Consequently, the sum

$$\sum_{k=1}^{n} |z - z_k|^2$$

is constant if and only if $n|z|^2 + nD_n^2$ is constant. The set of all such z is a circle centered at the centroid O, if we allow the empty set and a single point to be considered as degenerate cases of a circle.

The special case of Theorem 14.1 in which there are three points forming the vertices of a triangle can be found in [51; Corollary to Theorem 275].

The key to Theorem 14.1 is (14.2), which holds for any set of $n + 1$ points z_1, z_2, ..., z_n and z for which the first n points satisfy (14.1). If z_1, z_2, ..., z_n lie on a circle of radius r with center at their centroid O, (14.2) gives us

$$\sum_{k=1}^{n} |z - z_k|^2 = n(|z|^2 + r^2). \tag{14.3}$$

This holds, in particular, if z_1, z_2, ..., z_n are the vertices of a regular n-gon, or more generally if they are the vertices of a centrally symmetric polygon. If z also lies on the circle of radius r the sum in (14.3) reduces to

$$\sum_{k=1}^{n} |z - z_k|^2 = 2nr^2, \tag{14.4}$$

a generalization of the Pythagorean Theorem, which is the special case $n = 2$.

Another interesting special case occurs when z is one of the vertices. Then one term in the sum in (14.4) vanishes, and we obtain:

Theorem 14.2. *The sum of the squares of the $n-1$ segments drawn from one vertex of a regular n-gon to the remaining vertices is equal to $2nr^2$, where r is the radius of the circumscribing circle.*

We have already used Theorem 14.2 in Chapter 3 to calculate cycloidal areas without calculus. The same argument used to prove Theorem 14.1 (using dot products of vectors instead of complex numbers) also solves a corresponding locus problem in 3-space:

Given a finite set of fixed points in 3-space, what is the locus of a point moving in such a way that the sum of the squares of its distances from the fixed points is constant?

As expected, the answer is a direct extension of Theorem 14.1. The locus is a sphere whose center is at the centroid of the fixed points (if we allow the empty set and a single point as degenerate cases of a sphere).

The two locus problems provide motivation for the rest of this chapter, which discusses several remarkable formulas for sums of squares of distances in a finite-dimensional space, with applications to geometry. The concept of centroid described in Chapter 12 plays an important role.

Application to a minimizing problem.

Figure 14.1a illustrates a classical minimizing problem. Given a line L and two points A and B on the same side of L. Which point P on L minimizes the sum of distances $AP+BP$? It is well known that the solution is obtained by the reflection principle: At the minimizing point the two angles shown in Figure 14.1a are equal.

A different question for the same configuration is illustrated in Figure 14.1b:

Which point P on L minimizes the sum of squares of distances AP^2+BP^2?

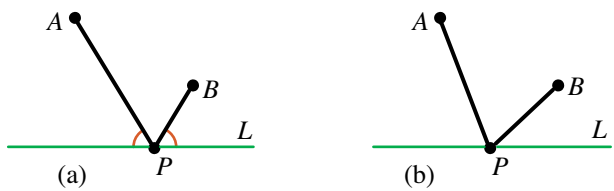


Figure 14.1: To minimize (a) sum of distances; (b) sum of squares of distances.

To answer this question we turn to the locus problem mentioned in the opening paragraph. The locus of all points in the plane for which the sum of squares is constant is a circle with its center C midway between A and B, and the constant is proportional to the square of the radius of the circle. Figure 14.2a shows a circle with center at C tangent to L. The point of tangency P minimizes the sum of squares, because all other points on L are outside the circle and therefore have larger sum of squares of distances from A and B. To construct P drop a perpendicular from

C to line L. Incidentally, for this problem A and B need not be on the same side of L.

A similar argument for minimizing the sum of distances $AP + BP$ is shown in Figure 14.2b. In this case the sum of distances is constant on an ellipse with foci at A and B, with its major axis proportional to the constant sum. Consequently, the point of tangency P of the ellipse with L minimizes $AP + BP$, and the known reflection property of the ellipse gives equality of the two angles in Figure 14.2b. By contrast, the corresponding angles in Figure 14.2a are not always equal. Instead, the projections of AP and BP on L are equal, as suggested by the tick marks.

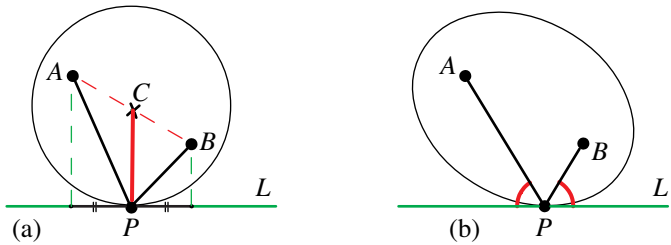


Figure 14.2: Minimizing (a) sum of squares of distances; (b) sum of distances.

The problem of minimizing the sum of squares can be extended in many ways. It is not necessary that A and B be coplanar with L. In fact, if we have n points in 3-space, the point P on a given line L that minimizes the sum of squares of the n distances from the points to L is obtained by dropping a perpendicular from the centroid C of the n points to that line. Moreover, the line L can be replaced by a space curve, a plane, or any surface. The normal from centroid C to L gives a local extremum of the sum of squares of distances. By contrast, the same question for the sum of distances is extremely difficult to analyze, even for three points.

14.2 FIRST BASIC THEOREM ON SUMS OF SQUARES OF DISTANCES IN m-SPACE

We begin with n arbitrary points $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$, regarded as position vectors in Euclidean m-space. Let $w_1, \ldots, w_n$, be n positive numbers regarded as weights attached to these points, and let $\boldsymbol{c}$ denote the weighted average position vector (or weighted centroid) defined by

$$\sum_{k=1}^{n} w_k \boldsymbol{r}_k = W\boldsymbol{c} \tag{14.5}$$

where

$$W = \sum_{k=1}^{n} w_k \tag{14.6}$$

is the sum of the weights. Then we have:

Theorem 14.3. *If $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$ are n points in m-space with weighted average $\boldsymbol{c}$, and if $\boldsymbol{z}$ is an arbitrary point, then*

$$\sum_{k=1}^{n} w_k|\boldsymbol{z}-\boldsymbol{r}_k|^2 = \sum_{k=1}^{n} w_k|\boldsymbol{c}-\boldsymbol{r}_k|^2 + W|\boldsymbol{c}-\boldsymbol{z}|^2. \tag{14.7}$$

Proof. We give a simple proof using dot products of vectors. The kth term of the sum on the left is

$$w_k|\boldsymbol{z}-\boldsymbol{r}_k|^2 = w_k(\boldsymbol{z}-\boldsymbol{r}_k)\cdot(\boldsymbol{z}-\boldsymbol{r}_k) = w_k(|\boldsymbol{z}|^2 + |\boldsymbol{r}_k|^2 - 2\boldsymbol{z}\cdot\boldsymbol{r}_k).$$

Similarly, the kth term of the sum on the right is

$$w_k|\boldsymbol{c}-\boldsymbol{r}_k|^2 = w_k(|\boldsymbol{c}|^2 + |\boldsymbol{r}_k|^2 - 2\boldsymbol{c}\cdot\boldsymbol{r}_k),$$

and their difference is

$$w_k|\boldsymbol{z}-\boldsymbol{r}_k|^2 - w_k|\boldsymbol{c}-\boldsymbol{r}_k|^2 = w_k(|\boldsymbol{z}|^2 - |\boldsymbol{c}|^2) + 2w_k(\boldsymbol{c}-\boldsymbol{z})\cdot\boldsymbol{r}_k\,.$$

Summing on k and using (14.5) and (14.6), we find that

$$\sum_{k=1}^{n} w_k|\boldsymbol{z}-\boldsymbol{r}_k|^2 - \sum_{k=1}^{n} w_k|\boldsymbol{c}-\boldsymbol{r}_k|^2 = W(|\boldsymbol{z}|^2 - |\boldsymbol{c}|^2) + 2W(\boldsymbol{c}-\boldsymbol{z})\cdot\boldsymbol{c}$$

$$= W(|\boldsymbol{z}|^2 + |\boldsymbol{c}|^2 - 2\boldsymbol{z}\cdot\boldsymbol{c}) = W|\boldsymbol{c}-\boldsymbol{z}|^2,$$

which gives (14.7).

When $m = 2$, Theorem 14.3 is related to a result in classical mechanics called Steiner's parallel-axis theorem, which states that the moment of inertia of a rigid body about an arbitrary axis is equal to the moment of inertia about a parallel axis through the center of mass of the body, plus the mass times the square of the distance between the two axes. Theorem 14.3 can be regarded as both a discrete analog of the parallel-axis theorem and an extension of it to higher dimensions.

An application to physics.

For an application of Theorem 14.3 to physics, let each point $\boldsymbol{r}_k$ be a center of attraction with a force acting on a particle $\boldsymbol{z}$ according to Hooke's law:

$$\boldsymbol{F}_k = -w_k(\boldsymbol{z}-\boldsymbol{r}_k). \tag{14.8}$$

This can be realized by an ideal spring or rubber band pulling particle $\boldsymbol{z}$ towards $\boldsymbol{r}_k$ with a force whose magnitude is proportional to the distance between them. In this context, the scalar w_k is called Hooke's coefficient, or the elasticity coefficient.

Now assume a particle $\boldsymbol{z}$ is attracted to each of n centers $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$ according to Hooke's law, each with its own Hooke's coefficient. The total potential energy $U(\boldsymbol{z})$ of particle $\boldsymbol{z}$ due to the attraction of all points $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$ can be defined by

$$U(\boldsymbol{z}) = \frac{1}{2}\sum_{k=1}^{n} w_k|\boldsymbol{z} - \boldsymbol{r}_k|^2. \tag{14.9}$$

The locus of points $\boldsymbol{z}$ for which $U(\boldsymbol{z})$ is a constant (independent of $\boldsymbol{z}$) is called an equipotential surface. The sum in (14.9) appears on the left of (14.7). The first term on the right of (14.7) does not depend on $\boldsymbol{z}$, hence $U(\boldsymbol{z})$ is constant if and only if the second term $W|\boldsymbol{c} - \boldsymbol{z}|^2$ is constant. Therefore, the equipotential surface is a sphere with center at the weighted centroid $\boldsymbol{c}$. This yields the following astonishing result: For any finite collection of fixed centers that attract by Hooke's law, the equipotential surfaces are spheres with center at the weighted centroid.

From (14.8) we find that the resultant force $\boldsymbol{F}(\boldsymbol{z})$ of the centers acting on a particle is

$$\boldsymbol{F}(\boldsymbol{z}) = \sum_{k=1}^{n} \boldsymbol{F}_k = -\sum_{k=1}^{n} w_k(\boldsymbol{z} - \boldsymbol{r}_k).$$

Replace the difference $\boldsymbol{z} - \boldsymbol{r}_k$ by $(\boldsymbol{z} - \boldsymbol{c}) + (\boldsymbol{c} - \boldsymbol{r}_k)$ and use (14.5) and (14.6) to obtain the simple formula $\boldsymbol{F}(\boldsymbol{z}) = -W(\boldsymbol{z} - \boldsymbol{c})$.

Thus the resultant is a central force directed toward $\boldsymbol{c}$, with its magnitude proportional to the distance of the particle from $\boldsymbol{c}$. In this case, it is known that the particle moves along an ellipse with its center at $\boldsymbol{c}$. Also, as shown in Section 1.16, its angular momentum is conserved.

The problem treated at the end of Section 14.1 on minimizing the sum of squares of distances can be realized physically as shown by the diagram in Figure 14.3a. Here two ideal rubber bands (subject to Hooke's law) join points A and P, and B and P, where P is attached to a ring surrounding a fixed horizontal rod L. The ring is free to slide along L, and at each position on L the ring is attracted to A and B by Hooke's law. The ring will reach an equilibrium point when the potential energy is a minimum. Because the potential energy is proportional to the sum of squares of distances from the ring to A and B, minimizing the potential energy is equivalent to minimizing the sum of squares of distances. As shown earlier, this occurs when the ring is at the point where the perpendicular through the centroid C intersects L.

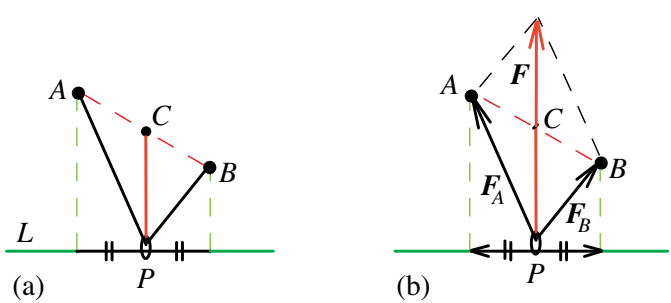


Figure 14.3: (a) Minimizing potential energy. (b) Equilibrium of an ideal rubber band.

An alternative solution is presented in Figure 14.3b, where two forces indicated by $\boldsymbol{F}_A$ and $\boldsymbol{F}_B$ act on P. By Hooke's law, their magnitudes are proportional to the lengths AP and BP. Their resultant $\boldsymbol{F}$ is determined by the parallelogram law. For equilibrium the resultant $\boldsymbol{F}$ must be perpendicular to L at P. Also, the horizontal components of $\boldsymbol{F}_A$ and $\boldsymbol{F}_B$ must be equal and opposite. Consequently, $\boldsymbol{F}$ passes through the point C midway between A and B, in agreement with the solution in Figure 14.3a.

Sums of the type in Theorem 14.3 also occur in other physics problems related to kinetic energy of a system of particles (where the vectors $\boldsymbol{r}_k$ represent velocities), and in angular momentum problems (where the $\boldsymbol{r}_k$ represent position vectors and the w_k are proportional to the mass times the angular speed of the particle).

They also occur in probability theory. If the weight w_k represents the probability that a random variable takes the value $\boldsymbol{r}_k$, then $\boldsymbol{c}$ is the mathematical expectation of the random variable, and $\sum_{k=1}^{n} w_k|\boldsymbol{c}-\boldsymbol{r}_k|^2$ is its variance (see [39]). Although most of the applications in this chapter are confined to geometry, the scope of possible applications is much larger.

14.3 SECOND BASIC THEOREM

The next result, a consequence of Theorem 14.3, provides another fundamental relation that is basic to all applications in this chapter. Before stating it we establish a notation convention. We use the symbol $\sum_{k<i}$ as an abbreviation for the double sum

$$\sum_{i=2}^{n}\sum_{k=1}^{i-1}.$$

Theorem 14.4. *Let $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$ be n points in m-space with weighted average $\boldsymbol{c}$. Then*

$$\sum_{k<i} w_i w_k|\boldsymbol{r}_i - \boldsymbol{r}_k|^2 = W\sum_{k=1}^{n} w_k|\boldsymbol{z} - \boldsymbol{r}_k|^2 - W^2|\boldsymbol{c} - \boldsymbol{z}|^2, \tag{14.10}$$

where $\boldsymbol{z}$ is arbitrary. In particular, when $\boldsymbol{z} = \boldsymbol{O}$ this becomes

$$\sum_{k<i} w_i w_k|\boldsymbol{r}_i - \boldsymbol{r}_k|^2 = W\sum_{k=1}^{n} w_k|\boldsymbol{r}_k|^2 - W^2|\boldsymbol{c}|^2, \tag{14.11}$$

and when $\boldsymbol{z} = \boldsymbol{c}$ it reduces to

$$\sum_{k<i} w_i w_k|\boldsymbol{r}_i - \boldsymbol{r}_k|^2 = W\sum_{k=1}^{n} w_k|\boldsymbol{c} - \boldsymbol{r}_k|^2. \tag{14.12}$$

Proof. Taking $\boldsymbol{z} = \boldsymbol{r}_i$ in (14.7), we obtain

$$\sum_{k=1}^{n} w_k|\boldsymbol{r}_i - \boldsymbol{r}_k|^2 = \sum_{k=1}^{n} w_k|\boldsymbol{c} - \boldsymbol{r}_k|^2 + W|\boldsymbol{c} - \boldsymbol{r}_i|^2$$

for $i = 1, 2, \ldots, n$ where, of course, the term with $k = i$ in the first sum is zero. Now multiply by w_i and sum on i to get

$$\sum_{i=1}^{n}\sum_{k=1}^{n} w_i w_k |\boldsymbol{r}_i - \boldsymbol{r}_k|^2 = W\sum_{k=1}^{n} w_k |\boldsymbol{c} - \boldsymbol{r}_k|^2 + W\sum_{i=1}^{n} w_i |\boldsymbol{c} - \boldsymbol{r}_i|^2$$

$$= 2W\sum_{k=1}^{n} w_k |\boldsymbol{c} - \boldsymbol{r}_k|^2.$$

This is equivalent to (14.12), because each term in the double sum appears twice. Now from (14.7) we have

$$W\sum_{k=1}^{n} w_k |\boldsymbol{c} - \boldsymbol{r}_k|^2 = W\sum_{k=1}^{n} w_k |\boldsymbol{z} - \boldsymbol{r}_k|^2 - W^2 |\boldsymbol{c} - \boldsymbol{z}|^2,$$

which, when substituted for the right-hand member of (14.12), gives (14.10).

Equal weights.

When all weights are equal, the weighted average $\boldsymbol{c}$ is the centroid of the set of points, and the foregoing relations connect sums of squares of distances among given points with squares of distances involving their centroid $\boldsymbol{c}$. If each $w_k = w$, say, then $W = nw$, and the common factor w cancels in each of the foregoing equations, which can now be written as indicated in equations (14.7′) through (14.12′).

From (14.7) we find that for any $\boldsymbol{z}$ in m-space,

$$\sum_{k=1}^{n} |\boldsymbol{z} - \boldsymbol{r}_k|^2 = \sum_{k=1}^{n} |\boldsymbol{c} - \boldsymbol{r}_k|^2 + n|\boldsymbol{c} - \boldsymbol{z}|^2. \tag{14.7′}$$

Specializing (14.10), (14.11), and (14.12) to equal weights we get:

$$\sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k|^2 = n\sum_{k=1}^{n} |\boldsymbol{z} - \boldsymbol{r}_k|^2 - n^2 |\boldsymbol{c} - \boldsymbol{z}|^2 \tag{14.10′}$$

for any $\boldsymbol{z}$ in m-space,

$$\sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k|^2 = n\sum_{k=1}^{n} |\boldsymbol{r}_k|^2 - n^2 |\boldsymbol{c}|^2, \tag{14.11′}$$

and

$$\sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k|^2 = n\sum_{k=1}^{n} |\boldsymbol{c} - \boldsymbol{r}_k|^2, \tag{14.12′}$$

which gives an exact relation connecting the sum of the squares of distances between arbitrary points $\boldsymbol{r}_1$, …, $\boldsymbol{r}_n$ in m-space with the sum of squares of their distances from the centroid.

For points in a plane, (14.11′) and (14.12′) can be found in Steiner [65, p. 108, (VI) and (VII)], and we thought their extensions to higher-dimensional space must surely be known. But our search of the literature did not uncover these exact formulas involving the centroid, even for three-dimensional space.

14.4 APPLICATIONS TO GEOMETRY

The importance of the foregoing results is revealed by a multitude of applications, some of which may be new.

Example 1 (Distances between arbitrary points). The exact relation in (14.11′) implies the inequality

$$\sum_{k=1}^{n} |\boldsymbol{r}_k|^2 \geq \frac{1}{n} \sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k|^2, \tag{14.13}$$

with equality if and only if the centroid $\boldsymbol{c}$ is at the origin. This tells us that the sum of squares of distances from a given point $\boldsymbol{O}$ to n fixed points $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$ is minimal when $\boldsymbol{O}$ is at their centroid, and the minimal value is the right-hand side of (14.13). This remarkable result was observed by Steiner for points in a plane ($m = 2$), but might not have been previously recorded for $m > 2$.

Example 2 (Points contained on a sphere in m-space). If $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$ and $\boldsymbol{z}$ lie on a sphere of radius R with center at the origin, and if $\boldsymbol{c} = \boldsymbol{O}$, (14.7′) reduces to

$$\sum_{k=1}^{n} |\boldsymbol{z} - \boldsymbol{r}_k|^2 = 2nR^2,$$

a generalization of the Pythagorean Theorem (which is the special case $n = 2$ when $\boldsymbol{r}_1 + \boldsymbol{r}_2 = \boldsymbol{O}$).

Now consider the case when all points $\boldsymbol{r}_k$ lie inside or on a sphere of radius R with center at the origin. Then $|\boldsymbol{r}_k| < R$ if $\boldsymbol{r}_k$ is inside the sphere, and $|\boldsymbol{r}_k| = R$ if $\boldsymbol{r}_k$ is on the sphere, so the sum on the left of (14.13) has the upper bound

$$\sum_{k=1}^{n} |\boldsymbol{r}_k|^2 \leq nR^2,$$

with equality if and only if all $\boldsymbol{r}_k$ are on the sphere. Equation (14.11′) now gives us:

Theorem 14.5. *For points $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$ in m-space with centroid $\boldsymbol{c}$ lying inside or on a sphere of radius R with center at the origin, the sum of the squares of distances $|\boldsymbol{r}_i - \boldsymbol{r}_k|$ satisfies the inequality*

$$\sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k|^2 \leq n^2(R^2 - |\boldsymbol{c}|^2), \tag{14.14}$$

with equality if and only if all points $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$ are on the sphere, in which case

$$\sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k|^2 = n^2(R^2 - |\boldsymbol{c}|^2). \tag{14.15}$$

This result provides a solution of Problem A4 of the 29th William Lowell Putnam Mathematical Competition [56], which asked to show that the sum of the squares of the $n(n-1)/2$ distances between a n distinct points on the surface of a unit sphere in 3-space is at most n^2. (See also [33] and 9.7 in [58], and our treatment in Section 12.7.) The explicit appearance of the centroid in (14.15) gives a deeper understanding of the problem. Equation (14.15) gives an exact formula for the sum of the squares of the distances between n points, no matter where they are located on a sphere in m-space, and it tells us that the sum reaches its maximum value n^2R^2 when the centroid is at the center of the sphere. This explains the surprising result that the maximum is independent of the dimensionality of the space. Any solution of the problem for n points lying on a circle is also a solution for n points lying on a sphere in m-space for all $m \geq 3$.

Example 3 (Sum of squares of edge lengths of a simplex). A simplex in m-space contains exactly $m+1$ vertices $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_{m+1}$ (which always lie on the circumscribed sphere), and exactly

$$\binom{m+1}{2} = \frac{m(m+1)}{2}$$

edges joining the vertices. Then $|\boldsymbol{r}_i - \boldsymbol{r}_k|$ is the length of the edge joining $\boldsymbol{r}_i$ and $\boldsymbol{r}_k$, and $|\boldsymbol{c} - \boldsymbol{r}_k|$ is the distance from the centroid $\boldsymbol{c}$ to vertex $\boldsymbol{r}_k$. Equation (14.12′) asserts that the sum of the squares of the lengths of the edges of a simplex is exactly $m+1$ times the sum of the squares of the distances from the centroid to the vertices. Inequality (14.13) now implies the inequality

$$\sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k|^2 \leq (m+1) \sum_{k=1}^{m+1} |\boldsymbol{r}_k|^2, \tag{14.16}$$

and tells us that the same sum is no more than $m+1$ times the sum of the squares of the distances from any point $\boldsymbol{O}$ to the vertices, with equality if and only if point $\boldsymbol{O}$ is the centroid.

In particular, (14.16) implies that in 2-space the sum of the squares of the lengths of the edges of any triangle is three times the sum of the squares of the distances from the centroid to the vertices, and no more than three times the sum of the squares of the distances from a point in the plane to the vertices. These properties of the triangle are known, but we were unable to locate the corresponding results in the literature for a tetrahedron or for any higher-dimensional simplex.

Example 4 (Sum of edge lengths of a simplex). The results in Example 3 for the sum of squares of the edge lengths of a simplex lead to upper bounds for the sum of the edge lengths themselves. The connection is provided by the Cauchy-Schwarz inequality, which states that

$$(\boldsymbol{u} \cdot \boldsymbol{v})^2 \leq |\boldsymbol{u}|^2 |\boldsymbol{v}|^2 \tag{14.17}$$

for any two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ in a finite-dimensional Euclidean space, with equality if and only if one of the vectors is a scalar multiple of the other. We use (14.15) to prove the following lemma concerning the sum of edge lengths of a simplex.

Lemma 1. *The square of the sum of the edge lengths of a simplex in m-space is less than or equal to the number of edges times the sum of squares of the lengths of its edges, with equality if and only if the simplex is regular.*

Proof. Denote the edge lengths of a simplex in m-space by $d_1, d_2, \ldots, d_N$, where $N = m(m+1)/2$ is the number of edges, and consider the following two vectors in N-space:

$$\boldsymbol{u} = (1, 1, \ldots, 1), \qquad \boldsymbol{v} = (d_1, d_2, \ldots, d_N).$$

Then (14.17) gives us

$$(\sum_{k=1}^{N} d_k)^2 \le N \sum_{k=1}^{N} d_k^2, \tag{14.18}$$

with equality if and only if $d_1 = d_2 = \cdots = d_N$.

Because $|\boldsymbol{r}_i - \boldsymbol{r}_k|$ is the length of the edge joining $\boldsymbol{r}_i$ and $\boldsymbol{r}_k$, (14.18) becomes

$$(\sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k|)^2 \le \frac{m(m+1)}{2} \sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k|^2, \tag{14.19}$$

with equality if and only if the simplex is regular.

In particular, when $m = 2$, this is a well-known inequality [58; p.147]. It states that the square of the perimeter of any triangle is less than or equal to three times the sum of the squares of the lengths of its edges, with equality if and only if the triangle is equilateral. Its three-dimensional counterpart says that the square of the sum of all the edge lengths of a tetrahedron is less than or equal to six times the sum of the squares of the lengths of its edges, with equality if and only if the tetrahedron is regular.

Taking square roots in (14.19), we are led to a basic inequality relating the sum of edge lengths of a simplex with the sum of squares of edge lengths:

$$\sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k| \le \sqrt{\frac{m(m+1)}{2} \sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k|^2}, \tag{14.20}$$

with equality if and only if the simplex is regular.

Using (14.12′) and (14.11′) for the sum of squares of edge lengths in (14.20), we discover corresponding results for the sum of edge lengths themselves, which we collect in the following theorem.

Theorem 14.6. *Let $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, …, $\boldsymbol{r}_{m+1}$ be the vertices of a simplex in m-space with centroid $\boldsymbol{c}$. Then the sum of edge lengths satisfies the two inequalities*

$$\sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k| \le (m+1) \sqrt{\frac{m}{2} \sum_{k=1}^{m+1} |\boldsymbol{c} - \boldsymbol{r}_k|^2}, \tag{14.21}$$

and

$$\sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k| \le (m+1) \sqrt{\frac{m}{2} \sum_{k=1}^{m+1} |\boldsymbol{r}_k|^2 - (m+1)|\boldsymbol{c}|^2}, \tag{14.22}$$

with equality in each case if and only if the simplex is regular and $\boldsymbol{c}$ *is at the origin. When* $\boldsymbol{c} = \boldsymbol{O}$ *they reduce to*

$$\sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k| \le (m+1)\sqrt{\frac{m}{2}\sum_{k=1}^{m+1} |\boldsymbol{r}_k|^2}, \tag{14.23}$$

with equality if and only if the simplex is regular.

Now suppose that each vertex of the simplex lies inside or on a sphere of radius R with center at the origin. Then we can use (14.14) from Theorem 14.5 (with $n = m+1$) in the right-hand side of (14.22) to obtain

$$\sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k| \le \sqrt{\frac{m(m+1)}{2}}(m+1)\sqrt{R^2 - |\boldsymbol{c}|^2}. \tag{14.24}$$

Equality holds in (14.24) if and only if the simplex is regular and has its vertices on the sphere, in which case $\boldsymbol{c} = \boldsymbol{O}$ and we get

$$\sum_{k<i} |\boldsymbol{r}_i - \boldsymbol{r}_k| \le \sqrt{\frac{m(m+1)}{2}}(m+1)R \tag{14.25}$$

for the sum of edge lengths. This agrees with a result found by Chakerian and Klamkin [33], who showed that among all simplices inscribed in the unit sphere, the regular simplex has maximum total edge length. If d denotes the common value of the edge lengths, the left-hand side of (14.25) is $dm(m+1)/2$, and with equality in (14.25) we find that

$$d = \sqrt{\frac{2(m+1)}{m}}R.$$

This expresses the length d of one edge of a regular simplex in m-space in terms of the radius R of the circumscribing sphere.

When $m = 2$, (14.25) gives $3\sqrt{3}R$ as the maximum perimeter of a triangle inscribed in a circle of radius R, and the maximum is achieved for an equilateral triangle, in which case $d = \sqrt{3}R$. The corresponding maximum sum of edge lengths for a tetrahedron in 3-space is $4\sqrt{6}R$, achieved for an inscribed regular tetrahedron of edge length $d = \sqrt{8/3}R$.

The formula for d enables us to find the angle between two radii drawn from the center of the tetrahedron to the endpoints of one edge. The result is $\pi - \arccos(1/3)$. The corresponding result in m-space is given by $\pi - \arccos(1/m)$. The distance d and the angle between two radii can also be calculated directly using elementary trigonometry but this requires much more effort. (Try it!)

14.5 COMPOSITE SYSTEMS

This section uses Theorem 14.3 to deduce a fundamental property of a concept we call the intrinsic second moment. If $w_1, w_2, \ldots, w_n$ are n positive weights attached to n points $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n$ in m-space with weighted average position vector $\boldsymbol{c}$ defined by (14.5), we refer to the quantity

$$I = \sum_{k=1}^{n} w_k |\boldsymbol{r}_k - \boldsymbol{c}|^2 \tag{14.26}$$

as the *intrinsic second moment* of the system.

A system of n_1 points $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_{n_1}$ in m-space with weights of sum W_1, average position vector $\boldsymbol{c}_1$, and intrinsic second moment I_1, taken together with a disjoint system of n_2 points $\boldsymbol{r}_{n_1+1}, \boldsymbol{r}_{n_1+2}, \ldots, \boldsymbol{r}_{n_1+n_2}$ with weights of sum W_2, average position vector $\boldsymbol{c}_2$, and intrinsic second moment I_2, forms a composite system consisting of $n_1 + n_2$ points with average position vector $\boldsymbol{c}$ and intrinsic second moment I. The intrinsic second moments are related as follows:

Theorem 14.7. *The intrinsic second moment of a composite system is related to the intrinsic second moments of the component systems by*

$$I = I_1 + I_2 + \frac{W_1 W_2}{W_1 + W_2} |\boldsymbol{c}_1 - \boldsymbol{c}_2|^2. \tag{14.27}$$

The quantity $W_1 W_2/(W_1 + W_2)$ is called the *reduced weight* of the composite system (by analogy with reduced mass in physics). Equation (14.27) states that the intrinsic second moment of a composite body is equal to the sum of the intrinsic second moments of the component parts plus the reduced weight times the square of the distance between their average position vectors. For $m = 2$, this is a known result for the moment of inertia of a composite body.

Proof. The definition in (14.26) with $n = n_1 + n_2$ gives us

$$I = \sum_{(1)} w_k |\boldsymbol{r}_k - \boldsymbol{c}|^2 + \sum_{(2)} w_k |\boldsymbol{r}_k - \boldsymbol{c}|^2,$$

where $\sum_{(1)}$ signifies a sum taken over the n_1 points in one system, and $\sum_{(2)}$ indicates summation over the n_2 points in the other. Applying Theorem 14.3 to each sum on the right with $\boldsymbol{z} = \boldsymbol{c}$, we obtain

$$\sum_{(1)} w_k |\boldsymbol{r}_k - \boldsymbol{c}|^2 = \sum_{(1)} w_k |\boldsymbol{r}_k - \boldsymbol{c_1}|^2 + W_1 |\boldsymbol{c}_1 - \boldsymbol{c}|^2,$$

and there is a corresponding formula for $\sum_{(2)}$. Adding them, we find

$$I = I_1 + I_2 + W_1 |\boldsymbol{c}_1 - \boldsymbol{c}|^2 + W_2 |\boldsymbol{c}_2 - \boldsymbol{c}|^2,$$

so to prove (14.27) it suffices to show that

$$W_1 |\boldsymbol{c}_1 - \boldsymbol{c}|^2 + W_2 |\boldsymbol{c}_2 - \boldsymbol{c}|^2 = \frac{W_1 W_2}{W_1 + W_2} |\boldsymbol{c}_1 - \boldsymbol{c}_2|^2. \tag{14.28}$$

From (14.5) we have $(W_1 + W_2)\boldsymbol{c} = W_1\boldsymbol{c}_1 + W_2\boldsymbol{c}_2$, hence the three averages $\boldsymbol{c}$, $\boldsymbol{c}_1$, and $\boldsymbol{c}_2$ are collinear, and $W_1(\boldsymbol{c} - \boldsymbol{c_1}) = W_2(\boldsymbol{c}_2 - \boldsymbol{c})$, which implies that

$$W_1|\boldsymbol{c} - \boldsymbol{c_1}| = W_2|\boldsymbol{c}_2 - \boldsymbol{c}| \tag{14.29}$$

and

$$|\boldsymbol{c} - \boldsymbol{c_1}| + |\boldsymbol{c}_2 - \boldsymbol{c}| = |\boldsymbol{c_1} - \boldsymbol{c}_2|. \tag{14.30}$$

Solving for $|\boldsymbol{c}_2 - \boldsymbol{c}|$ and substituting in (14.29), we obtain

$$(W_1 + W_2)|\boldsymbol{c} - \boldsymbol{c}_1| = W_2|\boldsymbol{c}_1 - \boldsymbol{c}_2|,$$

from which we conclude that

$$W_1|\boldsymbol{c} - \boldsymbol{c_1}| = \frac{W_1W_2}{W_1 + W_2}|\boldsymbol{c}_1 - \boldsymbol{c}_2|. \tag{14.31}$$

Now return to the left-hand member of (14.28) and rewrite it as:

$$(W_1|\boldsymbol{c}_1 - \boldsymbol{c}|)|\boldsymbol{c}_1 - \boldsymbol{c}| + (W_2|\boldsymbol{c}_2 - \boldsymbol{c}|)|\boldsymbol{c}_2 - \boldsymbol{c}|.$$

By (14.29) we can replace $W_2|\boldsymbol{c}_2 - \boldsymbol{c}|$ in the second term with $W_1|\boldsymbol{c}_1 - \boldsymbol{c}|$ and the sum becomes

$$(W_1|\boldsymbol{c}_1 - \boldsymbol{c}|)(|\boldsymbol{c}_1 - \boldsymbol{c}| + |\boldsymbol{c}_2 - \boldsymbol{c}|) =$$

$$(W_1|\boldsymbol{c}_1 - \boldsymbol{c}|)(|\boldsymbol{c}_1 - \boldsymbol{c_2}|) = \frac{W_1W_2}{W_1 + W_2}|\boldsymbol{c}_1 - \boldsymbol{c}_2|^2,$$

where we have used (14.30) and (14.31). This proves (14.28) and hence (14.27).

Relations not requiring the average position vector.

Although the intrinsic second moment is defined by (14.26) in terms of the average position vector $\boldsymbol{c}$, it can also be expressed in a form that does not involve $\boldsymbol{c}$, but only distances between points of the body. In fact, by using (14.12) of Theorem 14.4 in (14.26) we arrive at the expression

$$I = \frac{1}{W}\sum_{k<i} w_iw_k|\boldsymbol{r}_i - \boldsymbol{r}_k|^2, \tag{14.32}$$

and similar formulas hold for I_1 and I_2. Now multiply each side of (14.27) by $W = W_1 + W_2$ and rewrite it in the alternative form

$$\sum_{k<i} w_iw_k|\boldsymbol{r}_i - \boldsymbol{r}_k|^2 = \frac{W}{W_1}\sum_{(1)} + \frac{W}{W_2}\sum_{(2)} + W_1W_2|\boldsymbol{c}_1 - \boldsymbol{c}_2|^2, \tag{14.33}$$

where the sum on the left is extended over all points of the composite system, while $\sum_{(1)}$ and $\sum_{(2)}$ are the corresponding sums with the same summand $w_iw_k|\boldsymbol{r}_i - \boldsymbol{r}_k|^2$ but extended over the points of the respective component system. But we also have

$$\sum_{k<i} w_iw_k|\boldsymbol{r}_i - \boldsymbol{r}_k|^2 = \sum_{(1)} + \sum_{(2)} + \sum_{(1,2)},$$

where $\sum_{(1,2)}$ signifies summation with the same summand over vectors $\boldsymbol{r}_i - \boldsymbol{r}_k$ joining points from different component systems. Use this to replace the left-hand side of (14.33), and solve for $\sum_{(1,2)}$ to obtain

$$\sum_{(1,2)} = \frac{W_2}{W_1}\sum_{(1)} + \frac{W_1}{W_2}\sum_{(2)} + W_1 W_2 |\boldsymbol{c}_1 - \boldsymbol{c}_2|^2. \tag{14.34}$$

In other words, the double sum $\sum_{k<i} w_i w_k |\boldsymbol{r}_i - \boldsymbol{r}_k|^2$ extended over segments joining points $\boldsymbol{r}_i$ and $\boldsymbol{r}_k$ taken from two different component systems is a linear combination of sums extended over the separate systems plus the product of the weights times the square of the distance between the average position vectors of the two systems. This can be used to calculate the distance $|\boldsymbol{c}_1 - \boldsymbol{c}_2|$ without knowing $\boldsymbol{c_1}$ and $\boldsymbol{c}_2$ explicitly. And, if (14.34) is combined with (14.31), we can also compute the distance $|\boldsymbol{c} - \boldsymbol{c}_1|$ without knowing $\boldsymbol{c}$ or $\boldsymbol{c}_1$. In this case the formula for $|\boldsymbol{c} - \boldsymbol{c}_1|$ becomes

$$|\boldsymbol{c} - \boldsymbol{c}_1|^2 = \frac{1}{W^2 W_1^2}(W_1 W_2 \sum_{(1,2)} - W_2^2 \sum_{(1)} - W_1^2 \sum_{(2)}), \tag{14.35}$$

with summand $w_i w_k |\boldsymbol{r}_i - \boldsymbol{r}_k|^2$ in each sum on the right.

14.6 EQUAL WEIGHTS: APPLICATIONS TO GEOMETRY

This section considers composite systems having equal weights, with the total weight of each system replaced by the corresponding number of points n_1, n_2 and $n = n_1 + n_2$. Formulas (14.33) and (14.37) of Section 14.5 have interesting applications to geometry.

Example 5 (General simplex). An m-dimensional simplex with $m+1$ vertices can be regarded as a composite system made up of two parts, one part having exactly one vertex, which is also its centroid $\boldsymbol{c}_1$, and the other part consisting of the remaining m vertices. Using (14.35) to determine the distance from $\boldsymbol{c}_1$ to the centroid $\boldsymbol{c}$ of the entire simplex, we find that

$$|\boldsymbol{c} - \boldsymbol{c}_1| = \frac{1}{m+1}\sqrt{m\sum_{(1,2)} - \sum_{(2)}},$$

a result consistent with (14.7′) when $n = m + 1$ and $\boldsymbol{z} = \boldsymbol{c}_1$. In this case, $\sum_{(1,2)}$ is the sum of squares of distances from $\boldsymbol{c}_1$ to all adjacent vertices, $\sum_{(1)}$ is zero, and $\sum_{(2)}$ is the sum of squares of all the remaining edges. For example, for a triangle ($m = 2$) with edges of lengths a_1, a_2, and a_3, where a_1 and a_2 are the lengths of the edges adjacent to $\boldsymbol{c}_1$, we have

$$|\boldsymbol{c} - \boldsymbol{c}_1| = \frac{1}{3}\sqrt{2(a_1^2 + a_2^2) - a_3^2}.$$

The corresponding formula for a tetrahedron ($m = 3$) is

$$|\boldsymbol{c} - \boldsymbol{c}_1| = \frac{1}{4}\sqrt{3(a_1^2 + a_2^2 + a_3^2) - (a_4^2 + a_5^2 + a_6^2)}$$

where a_1, a_2, and a_3 are the lengths of the edges adjacent to $\boldsymbol{c}_1$ and a_4, a_5, and a_6 are the lengths of the remaining edges.

Example 6 (Quadrilaterals). Let $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, $\boldsymbol{r}_3$ and $\boldsymbol{r}_4$ be four distinct points in m-space. Take $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ as one system, and $\boldsymbol{r}_3$ and $\boldsymbol{r}_4$ as another. Use equal weights, let $\boldsymbol{c}_1$ be the midpoint of the segment joining $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$, and let $\boldsymbol{c}_2$ be the midpoint of the segment joining $\boldsymbol{r}_3$ and $\boldsymbol{r}_4$. Equation (14.34) now becomes

$$|\boldsymbol{r}_1 - \boldsymbol{r}_3|^2 + |\boldsymbol{r}_1 - \boldsymbol{r}_4|^2 + |\boldsymbol{r}_2 - \boldsymbol{r}_3|^2 + |\boldsymbol{r}_2 - \boldsymbol{r}_4|^2$$

$$= |\boldsymbol{r}_1 - \boldsymbol{r}_2|^2 + |\,\boldsymbol{r}_3 - \boldsymbol{r}_4|^2 + 4\,|\boldsymbol{c}_1 - \boldsymbol{c}_2|^2. \tag{14.36}$$

The points $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, $\boldsymbol{r}_3$, $\boldsymbol{r}_4$ can be regarded as the vertices of a quadrilateral, with the sum on the left being the sum of the squares of the lengths of the four edges, and the first two terms on the right being the sum of the squares of the two diagonals. Equation (14.36) states that the sum of the squares of the edges exceeds the sum of the squares of the diagonals by four times the square of the distance between the midpoints of the diagonals. This extends a well-known planar result to m-space. It also shows that the sum of the squares of the diagonals is at most the sum of the squares of its four edges, with equality if and only if $\boldsymbol{c_1} = \boldsymbol{c}_2$. The latter occurs only when the diagonals bisect each other, that is, when the quadrilateral is a parallelogram lying in a plane. Example 7 gives a corresponding result for hexagons.

Tetrahedron problem. Figure 14.4a shows an application of (14.36) to a problem involving a general tetrahedron. Take an arbitrary tetrahedron in 3-space whose six edge lengths are known. The problem is to find the length of the segment joining midpoints of a pair of opposite sides.

Solution. In Figure 14.4a, the edges have lengths a, b, c, d, e, f with e and f representing one pair of opposite edges. By applying (14.36) we find that the length of the segment joining their midpoints is equal to

$$\frac{1}{2}\sqrt{a^2 + b^2 + c^2 + d^2 - e^2 - f^2},$$

regardless of how the other four edges are ordered.

Equilibrium problem. Figure 14.4b shows two rigid rods AB and CD in 3-space with their endpoints connected by ideal rubber bands as indicated. Now the problem is to determine the equilibrium configuration of the rods.

Solution. The two rods form the diagonals and the four rubber bands form the edges of a quadrilateral satisfying (14.36). Because potential energy is proportional to the sum of squares of the edges, its minimum will be reached when the centroids of the two rods coincide. This will provide the equilibrium configuration in the form of a parallelogram with the rods as diagonals bisecting each other. The parallelogram is not unique: the minimum potential energy is the same for all possible parallelograms regardless of the angle between the diagonal rods. This is a surprising example of what is known as indifferent equilibrium.

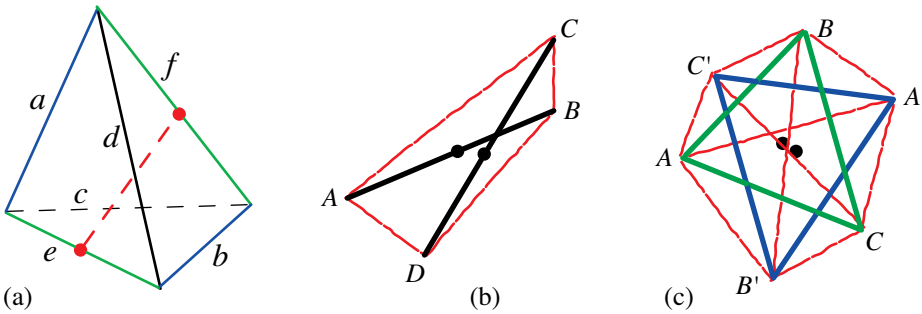


Figure 14.4: (a) Tetrahedron problem. (b) and (c): Equilibrium problems.

Example 7 (Hexagons). Let $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, …, $\boldsymbol{r}_6$ be six distinct points in m-space joined in sequence to form a hexagon. Vertices $\boldsymbol{r}_1$, $\boldsymbol{r}_3$, and $\boldsymbol{r}_5$ form one triangle with centroid $\boldsymbol{c}_1$, and vertices $\boldsymbol{r}_2$, $\boldsymbol{r}_4$, and $\boldsymbol{r}_6$ form another triangle with centroid $\boldsymbol{c}_2$. In general, the triangles are in different planes, and they may or may not intersect. The hexagon has six minor diagonals that form the edges of the two triangles, plus three major diagonals joining $\boldsymbol{r}_1$ and $\boldsymbol{r}_4$, $\boldsymbol{r}_2$ and $\boldsymbol{r}_5$, and $\boldsymbol{r}_3$ and $\boldsymbol{r}_6$, respectively. Take equal weights and use the abbreviation r_{ij}^2 for $|\boldsymbol{r}_i - \boldsymbol{r}_j|^2$, and (14.34) becomes

$$(r_{12}^2 + r_{23}^2 + r_{34}^2 + r_{45}^2 + r_{56}^2 + r_{61}^2) + (r_{14}^2 + r_{25}^2 + r_{36}^2)$$

$$= (r_{13}^2 + r_{35}^2 + r_{51}^2 + r_{24}^2 + r_{46}^2 + r_{62}^2) + 9|\boldsymbol{c}_1 - \boldsymbol{c}_2|^2.$$

This states that the sum of squares of the six edges and the three major diagonals of a hexagon exceeds the sum of the squares of its six minor diagonals by nine times the square of the distance between the centroids of the vertices of the two triangles. Equality holds if and only if the two centroids coincide, although the hexagon need not in that case be flat.

Equilibrium problem. Figure 14.4c shows two triangles ABC and $A'B'C'$, not necessarily coplanar, whose edges are rigid rods. The non-connected vertices are joined by ideal rubber bands forming a hexagon and its major diagonals as shown. The rigid edges are minor diagonals of the hexagon. The problem is to determine the equilibrium configuration of this structure.

Solution. In the last displayed formula, the terms on the left represent the sum of squares of the edges of the hexagon and its major diagonals. It reaches its minimum when the centroids of the vertices of the two triangles coincide, because the sum of squares of minor diagonals on the right is fixed. This gives another example of indifferent equilibrium, independent of the relative spatial orientation of the rigid triangles, as long as their centroids coincide. This is even more surprising than the parallelogram equilibrium in the quadrilateral case.

14.7 SUMS OF SQUARES OF INTEGERS IN ARITHMETIC PROGRESSION

The following striking identities

$$3^2 + 4^2 = 5^2$$
$$10^2 + 11^2 + 12^2 = 13^2 + 14^2$$
$$21^2 + 22^2 + 23^2 + 24^2 = 25^2 + 26^2 + 27^2$$
$$36^2 + 37^2 + 38^2 + 39^2 + 40^2 = 41^2 + 42^2 + 43^2 + 44^2$$

are the cases $n = 1, 2, 3, 4$ of a remarkable family given in a note by G. J. Dostor [38]. Dostor's family is

$$\sum_{k=0}^{n-1}(k+m)^2 + (m+n)^2 = \sum_{k=0}^{n-1}(k+m+n+1)^2, \tag{14.37}$$

where $m = n(2n+1)$, and $n = 1, 2, \ldots$. We assume that $m \neq -n$, because $m = -n$ implies $n = 1$, a trivial case. Then there are $n+1$ squares of consecutive integers on the left, and n on the right. We will treat the last term $(m+n)^2$ on the left differently, and refer to it as a *transition term* relating two sums of squares of n consecutive integers.

This section generalizes (14.37) in several simple ways, including extensions to arithmetic progressions. All are based on the elementary factorization formula

$$x^2 - y^2 = (x-y)(x+y).$$

First generalization of (14.37): change of transition term.

Given an integer $r \geq 1$, we show that integer n and rational m exist (depending on r) satisfying the family of identities

$$\sum_{k=0}^{n-1}(k+m)^2 + (m+n)^2 r^2 = \sum_{k=0}^{n-1}(k+m+n+1)^2. \tag{14.38}$$

This is like Dostor's family, except that the new transition term has a factor r^2, and m is a new parameter.

To see that m and n exist, take the difference of the sums in (14.38) and factor $(k+m+n+1)^2 - (k+m)^2$ to obtain

$$\sum_{k=0}^{n-1}(n+1)[2(k+m)+n+1] = 2n(n+1)(m+n).$$

This equals $(m+n)^2 r^2$ if and only if $2n(n+1) = r^2(m+n)$, which is equivalent to

$$m = \frac{2n(n+1)}{r^2} - n. \tag{14.39}$$

Therefore, (14.38) is valid for every choice of positive integers r and n and rational m satisfying (14.39). In (14.38), there are n consecutive squares on the left, plus a transition term $(m+n)^2r^2$, and n consecutive squares on the right.

If $m = a/b$ where a and b are integers, then (14.38) can be written as

$$\sum_{k=0}^{n-1}(bk+a)^2 + (bn+a)^2r^2 = \sum_{k=n+1}^{2n}(bk+a)^2, \;\; n = 1, 2, 3, \ldots, \tag{14.40}$$

where now each sum contains squares of n integers in arithmetic progression. Because the number $bn + a$ in the transition term is midway between $b(n-1)+a$ and $b(n+1)+a$, we say that (14.40) has a *symmetry property* with respect to the transition term.

When $b = 1$, then $a = m$, an integer, and (14.40) gives (14.38) with consecutive squares starting with m^2 on the left. From (14.39) we see that m is an integer whenever r^2 divides the product $2n(n+1)$. So we can always choose $n(n+1)$ to be a multiple of r^2 and find corresponding integer values of m.

Examples with $r = 1$. If $r = 1$, then $m = n(2n+1)$, and family (14.38) is the same as Dostor's family (14.37).

Examples with $r = 2$. If $r = 2$, then $m = n(n-1)/2$, which is always an integer. The case $n = 1$ is trivial, and $n = 2, 3, 4$, yield

$$1^2 + 2^2 + 3^2 \cdot 2^2 = 4^2 + 5^2$$
$$3^2 + 4^2 + 5^2 + 6^2 \cdot 2^2 = 7^2 + 8^2 + 9^2$$
$$6^2 + 7^2 + 8^2 + 9^2 + 10^2 \cdot 2^2 = 11^2 + 12^2 + 13^2 + 14^2.$$

Examples with $r = 3$. If $r = 3$, integer solutions m exist in (14.39) if $n = 9N - 1$ or if $n = 9N$ for some integer N. The two smallest positive n with $n = 9N - 1$ are $n = 8, m = 8$, and $n = 17$, $m = 51$, yielding:

$$8^2 + 9^2 + \cdots + 15^2 + 16^2 \cdot 3^2 = 17^2 + 18^2 + \cdots + 24^2$$
$$51^2 + 52^2 + \cdots + 67^2 + 68^2 \cdot 3^2 = 69^2 + 70^2 + \cdots + 85^2.$$

The two smallest with $n = 9N$ are $n = 9$, $m = 11$, and $n = 18$, $m = 58$:

$$11^2 + 12^2 + \cdots + 19^2 + 20^2 \cdot 3^2 = 21^2 + 22^2 + \cdots + 29^2$$
$$58^2 + 59^2 + \cdots + 75^2 + 76^2 \cdot 3^2 = 77^2 + 78^2 + \cdots + 94^2.$$

If $n = 2$ and 3, the values of m from (14.39) are $m = -2/3$ and $m = -1/3$, giving the following two identities involving squares of integers in arithmetic progression:

$$(-2)^2 + 1^2 + 4^2 \cdot 3^2 = 7^2 + 10^2$$
$$(-1)^2 + 2^2 + 5^2 + 8^2 \cdot 3^2 = 11^2 + 14^2 + 17^2.$$

If $n = 4, 5$, and 6 the values of m from (14.39) are $m = 4/9, m = 5/3$, and $m = 10/3$ and we get the identities

$$4^2 + 13^2 + 22^2 + 31^2 + 40^2 \cdot 3^2 = 49^2 + 58^2 + 67^2 + 76^2$$

$$5^2 + 8^2 + 11^2 + 14^2 + 17^2 + 20^2 \cdot 3^2 = 23^2 + 26^2 + 29^2 + 32^2 + 35^2$$

$$10^2 + 13^2 + 16^2 + 19^2 + 22^2 + 25^2 + 28^2 \cdot 3^2 = 31^2 + 34^2 + 37^2 + 40^2 + 43^2 + 46^2.$$

The foregoing examples show how easy it is to produce families from (14.38) or (14.40) for each integer r. All the identities exhibit the symmetry property with respect to the transition term.

Second generalization of (14.37): introduction of a shift parameter.

We introduce a new parameter d. For an integer n and real m and d we have the family of identities

$$\sum_{k=0}^{n-1}(k+m)^2 + nd(2c+d) = \sum_{k=0}^{n-1}(k+m+d)^2, \tag{14.41}$$

where c is the arithmetic mean:

$$c = \frac{1}{n}\sum_{k=0}^{n-1}(k+m). \tag{14.42}$$

The proof of (14.41) is obtained by summing both members of the factorization formula

$$(k+m+d)^2 - (k+m)^2 = d(2k+2m+d).$$

When m and d are integers, both sums in (14.41) involve squares of n consecutive integers, those on the right being shifted by d. In this case, the transition term $nd(2c+d)$ is also an integer because it is the difference of two integers. We are interested only in examples where the transition term is also a square, although it may not be consecutive with the other squares in the identity.

If m is rational, say $m = a/b$, then (14.41) becomes

$$\sum_{k=0}^{n-1}(bk+a)^2 + b^2nd(2c+d) = \sum_{k=0}^{n-1}(bk+a+bd)^2. \tag{14.43}$$

The terms $bk+a$ on the left are in arithmetic progression, and those on the right are shifted by bd. The identity is of interest when the transition term is the square of an integer, which again occurs when $nd(2c+d)$ is a square. To find examples of (14.43) for which $nd(2c+d)$ is a square, we express the arithmetic mean c in terms of the parameters m and n. From (14.42) we find that $2c = n - 1 + 2m$. If $g = d - n$, then $d = n + g$, and the transition term in (14.41) takes the form

$$nd(2c+d) = 2n(n+g)(n+m+\frac{g-1}{2}).$$

Symmetric families; symmetry requirement for the transition term.

In (14.41), the average of $n-1+m$ and $m+d$ is $(n+m+\frac{g-1}{2})$, and we get the symmetry property with respect to the transition term if and only if

$$2n(n+g)(n+m+\frac{g-1}{2}) = (n+m+\frac{g-1}{2})^2 r^2$$

for some integer $r \geq 1$. This is equivalent to

$$(n+m+\frac{g-1}{2})r^2 = 2n(n+g), \tag{14.44}$$

or

$$m = \frac{2n(n+g)}{r^2} - n - \frac{g-1}{2}. \tag{14.45}$$

We are interested in positive integer values of the parameters n, r, and g. The number m determined by (14.45) can be positive, negative, or 0. If $m = a/b$, where a and b are integers with $b > 0$, (14.41) becomes the following identity with squares of integers in arithmetic progression in each sum:

$$\sum_{k=0}^{n-1}(bk+a)^2 + b^2(n+m+\frac{g-1}{2})^2 r^2 = \sum_{k=0}^{n-1}(bk+a+bn+bg)^2. \tag{14.46}$$

When g is odd, the transition term is also the square of an integer. In particular, if g is odd and $b = 1$, then $a = m$ and each sum in (14.46) contains squares of n consecutive integers,

$$\sum_{k=0}^{n-1}(k+m)^2 + (n+m+\frac{g-1}{2})^2 r^2 = \sum_{k=n}^{2n-1}(k+m+g)^2, \tag{14.47}$$

with the terms shifted symmetrically with respect to the transition term.

Examples with $g = 1$. In this case, (14.45) reduces to (14.39), which gives (14.38) and (14.40) for any integer $r \geq 1$.

Examples with $g = 3$. Now (14.45) becomes

$$m = \frac{2n(n+3)}{r^2} - n - 1.$$

If $g = 3$ and $r = 1$, then $m = 2n^2 + 5n - 1$, and family (14.47) takes the form

$$\sum_{k=0}^{n-1}(k+2n^2+5n-1)^2 + (2n^2+6n)^2 = \sum_{k=n}^{2n-1}(k+2n^2+5n+2)^2. \tag{14.48}$$

The special cases $n = 1, 2, 3$ yield

$$6^2 + 8^2 = 10^2$$

$$17^2 + 18^2 + 20^2 = 22^2 + 23^2$$

$$32^2 + 33^2 + 34^2 + 36^2 = 38^2 + 39^2 + 40^2.$$

If $g = 3$ and $r = 2$, then $m = \frac{(n-1)(n+2)}{2}$ is an integer, and family (14.47) takes the form

$$\sum_{k=0}^{n-1}(k + \frac{(n-1)(n+2)}{2})^2 + (n(n+3))^2 = \sum_{k=n}^{2n-1}(k + \frac{n(n+1)}{2} + 2)^2. \qquad (14.49)$$

The case $n = 1$ is not interesting, and the special cases $n = 2, 3, 4$ yield

$$2^2 + 3^2 + 5^2 \cdot 2^2 = 7^2 + 8^2$$

$$5^2 + 6^2 + 7^2 + 9^2 \cdot 2^2 = 11^2 + 12^2 + 13^2$$

$$9^2 + 10^2 + 11^2 + 12^2 + 14^2 \cdot 2^2 = 16^2 + 17^2 + 18^2 + 19^2.$$

Families with $n = r^2/2$. In this case we take r even, say $r = 2N$, to guarantee an integer value for n. Then (14.45) implies $m = (g + 1)/2$. If g is odd, m is an integer, and if g is even, m is a rational a/b with $a = g + 1$ and $b = 2$. Therefore we separate the cases g odd, and g even.

(a) g odd, $m = (g + 1)/2$. Family (14.47) now has the transition term given by $(n + g)^2 r^2$, where $r = 2N, n = 2N^2$:

$$\sum_{k=0}^{2N^2-1}(k + m)^2 + (2N^2 + g)^2 \cdot (2N)^2 = \sum_{k=2N^2}^{4N^2-1}(k + m + g)^2. \qquad (14.50)$$

The smallest N is 1, giving $r = 2, n = 2$, and (14.50) takes the form

$$\sum_{k=0}^{1}(k + m)^2 + (2 + g)^2 \cdot 2^2 = \sum_{k=2}^{3}(k + m + g)^2.$$

The values $g = 1, 3, 5$ yield

$$1^2 + 2^2 + 3^2 \cdot 2^2 = 4^2 + 5^2$$

$$2^2 + 3^2 + 5^2 \cdot 2^2 = 7^2 + 8^2$$

$$3^2 + 4^2 + 7^2 \cdot 2^2 = 10^2 + 11^2.$$

More generally, when $N \geq 2$ and $g = 1$, then $m = 1$, and (14.50) takes the form

$$\sum_{k=1}^{2N^2} k^2 + (2N^2 + 1)^2(2N)^2 = \sum_{k=2N+2}^{4N^2+1} k^2. \qquad (14.51)$$

The values $N = 2, 3, 4$ and $g = 1$ yield

$$1^2 + 2^2 + 3^2 + \cdots + 7^2 + 8^2 + 9^2 \cdot 4^2 = 10^2 + 11^2 + 12^2 + \cdots + 16^2 + 17^2$$

$$1^2 + 2^2 + 3^2 + \cdots + 17^2 + 18^2 + 19^2 \cdot 6^2 = 20^2 + 21^2 + 22^2 + \cdots + 36^2 + 37^2$$

$$1^2 + 2^2 + 3^2 + \cdots + 31^2 + 32^2 + 33^2 \cdot 8^2 = 34^2 + 35^2 + 36^2 + \cdots + 64^2 + 65^2.$$

When $g = 3$, then $m = 2$, and (14.48) can be written as

$$\sum_{k=2}^{2N^2+1} k^2 + (2N^2+3)^2 \cdot (2N)^2 = \sum_{k=2N^2+5}^{4N^2+4} k^2, \tag{14.52}$$

which, for $N = 1, 2, 3$, yield the examples

$$2^2 + 3^2 + 5^2 \cdot 2^2 = 7^2 + 8^2$$

$$2^2 + 3^2 + 4^2 + \cdots + 8^2 + 9^2 + 11^2 \cdot 4^2 = 13^2 + 14^2 + 15^2 + \cdots + 19^2 + 20^2$$

$$2^2 + 3^2 + 4^2 + \cdots + 18^2 + 19^2 + 21^2 \cdot 6^2 = 23^2 + 24^2 + 25^2 + \cdots + 39^2 + 40^2.$$

(b) g even, $m = (g+1)/2$. The corresponding family contains terms in arithmetic progression as in (14.46), with $a = g+1, b = 2$, and with transition term $4(n+g)^2r^2$. Family (14.46) takes the form

$$\sum_{k=0}^{n-1}(2k+1+g)^2 + (2n+2g)^2r^2 = \sum_{k=0}^{n-1}(2k+1+g+2(n+g))^2, \tag{14.53}$$

with $r = 2N, n = 2N^2$. If $N = 1$, then $r = n = 2$, and (14.53) reduces to

$$\sum_{k=0}^{1}(2k+1+g)^2 + (4+2g)^2 \cdot 2^2 = \sum_{k=0}^{1}(2k+1+g+2(2+g))^2.$$

The values $g = 0, 2, 4$ yield

$$1^2 + 3^2 + 4^2 \cdot 2^2 = 5^2 + 7^2$$

$$3^2 + 5^2 + 8^2 \cdot 2^2 = 11^2 + 13^2$$

$$5^2 + 7^2 + 12^2 \cdot 2^2 = 17^2 + 19^2.$$

Families with $n + g = r^2/2$. In this case, (14.45) implies $m = (1-g)/2$. If g is odd, m is an integer; if g is even, $m = a/b$, with $a = 1-g$ and $b = 2$. For example, if $g = 1$, then $m = 0$, and the family in (14.47) has transition term n^2r^2, where n is chosen so that $2(n+1) = r^2$. Family (14.47) becomes

$$\sum_{k=0}^{n-1} k^2 + n^2r^2 = \sum_{k=n}^{2n-1}(k+1)^2. \tag{14.54}$$

The values $r = 4, 6, 8$, with $n = 7, 17, 31$, yield:

$$1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2 + 7^2 \cdot 4^2 = 8^2 + 9^2 + 10^2 + 11^2 + 12^2 + 13^2 + 14^2$$

$$1^2 + 2^2 + 3^2 + \cdots + 15^2 + 16^2 + 17^2 \cdot 6^2 = 18^2 + 19^2 + 20^2 + \cdots + 32^2 + 33^2 + 34^2$$

$$1^2 + 2^2 + 3^2 + \cdots + 29^2 + 30^2 + 31^2 \cdot 8^2 = 32^2 + 33^2 + 34^2 + \cdots + 60^2 + 61^2 + 62^2.$$

If $g = 0$, then $m = 1/2$ and (14.46) becomes

$$\sum_{k=0}^{n-1} (2k+1)^2 + (2n)^2 r^2 = \sum_{k=0}^{n-1} (2k + 1 + 2n)^2,$$

with $r^2 = 2n$. The values $r = 2, 4, 6$, with $n = 2, 8, 18$, yield

$$1^2 + 3^2 + 4^2 \cdot 2^2 = 5^2 + 7^2$$

$$1^2 + 3^2 + 5^2 + \cdots + 13^2 + 15^2 + 16^2 \cdot 4^2 = 17^2 + 19^2 + 21^2 + \cdots + 29^2 + 31^2$$

$$1^2 + 3^2 + 5^2 + \cdots + 33^2 + 35^2 + 36^2 \cdot 6^2 = 37^2 + 39^2 + 41^2 + \cdots + 69^2 + 71^2.$$

Controlling the first term.

The parameter m represents the first term in basic identity (14.41), and it is important to note that families exist where it is specified in advance. We can always choose m first and then determine g by the relations $g = 2m - 1$ if $m > 0$, or $g = 1 - 2m$ if $m < 0$. The foregoing requirements $n = r^2/2$ for $m > 0$, and $n + g = r^2/2$ for $m < 0$, relate parameters r and n. The following examples arise from the choices indicated.

The choice $m = 7$ gives $g = 13$. When $r = n = 2$, we get from (14.50) with $N = 1$:

$$7^2 + 8^2 + 15^2 \cdot 2^2 = 22^2 + 23^2,$$

and when $r = 4, n = 8$, we get from (14.50) with $N = 2$:

$$7^2 + 8^2 + 9^2 + \cdots + 13^2 + 14^2 + 21^2 \cdot 4^2 = 28^2 + 29^2 + 30^2 + \cdots + 34^2 + 35^2.$$

The choice $m = -2$ gives $g = 5$, and we use (14.47). When $r = 4$ and $n = 3$, we find

$$(-2)^2 + (-1)^2 + 0^2 + 3^2 \cdot 4^2 = 6^2 + 7^2 + 8^2,$$

and when $r = 6$, $n = 13$, we obtain

$$(-2)^2 + (-1)^2 + 0^2 + 1^2 + \cdots + 9^2 + 10^2 + 13^2 \cdot 6^2 = 16^2 + 17^2 + 18^2 + \cdots + 27^2 + 28^2.$$

Pythagorean triples.

When $n = 1$, (14.40) becomes $a^2 + (a+b)^2 r^2 = (a+2b)^2$, so $(a, (a+b)r, a+2b)$ is a Pythagorean triple if $a > 0$. Euclid's formulas $x = (u^2 - v^2), y = 2uv, z = (u^2 + v^2)$ for generating all Pythagorean triples correspond to (14.40) by taking $a = (u^2 - v^2), b = v^2$, and $r = 2v/u$, a rational multiplier. Because Euclid's formulas generate all Pythagorean triples, so does (14.40). If the multiplier r is not an integer, the resulting Pythagorean identity will not be symmetric with respect to the transition term. We turn now to asymmetric identities.

Asymmetric families.

The foregoing examples do not exhaust all possible families of identities. We have treated those with a symmetry property relative to the transition term as required by (14.44). There are many families that do not satisfy (14.44), which we call *asymmetric*. For example

$$1^2 + 2^2 + 3^2 + 4^2 + 5^2 \cdot 3^2 = 5^2 + 6^2 + 7^2 + 8^2 + 9^2$$

$$1^2 + 2^2 + 3^2 + \cdots + 11^2 + 12^2 + 13^2 \cdot 5^2 = 13^2 + 14^2 + 15^2 + \cdots + 24^2 + 25^2$$

$$1^2 + 2^2 + 3^2 + \cdots + 23^2 + 24^2 + 25^2 \cdot 7^2 = 25^2 + 26^2 + 27^2 + \cdots + 48^2 + 49^2$$

are members of the family

$$\sum_{k=1}^{2N} k^2 + (2N+1)^2 M^2 = \sum_{k=0}^{2N} (k + 2N + 1)^2, \tag{14.55}$$

in which $4N + 1 = M^2$. In the examples, $N = 2, 6, 12$, and $M = 3, 5, 7$.

Related examples are

$$1^2 + 2^2 + 3^2 + 4^2 + 4^2 \cdot 3^2 = 5^2 + 6^2 + 7^2 + 8^2$$

$$1^2 + 2^2 + 3^2 + \cdots + 11^2 + 12^2 + 12^2 \cdot 5^2 = 13^2 + 14^2 + 15^2 + \cdots + 24^2$$

$$1^2 + 2^2 + 3^2 + \cdots + 23^2 + 24^2 + 24^2 \cdot 7^2 = 25^2 + 26^2 + 27^2 + \cdots + 47^2 + 48^2,$$

which are members of the family:

$$\sum_{k=1}^{2N} k^2 + 4N^2 M^2 = \sum_{k=0}^{2N-1} (k + 2N + 1)^2, \tag{14.56}$$

in which $4N + 1 = M^2$. In the examples, $N = 2, 6, 12$, and $M = 3, 5, 7$.

Now we show that the asymmetric families can be obtained by returning to the basic identity (14.41) in which the transition term is $nd(2c + d)$, where c is given by (14.42), and making simple choices of parameters m and d.

Take n odd, say $n = 2N + 1$, choose $m = 0$, and $d = n = 2N + 1$. Then $c = (2N + 1)/2$ and (14.41) becomes

$$\sum_{k=1}^{2N} k^2 + (2N+1)^2(4N+1) = \sum_{k=0}^{2N} (k + 2N + 1)^2,$$

which gives (14.55) by requiring that $4N + 1$ be a square.

To obtain (14.56) from (14.55), transpose the last term on the right and observe that

$$(2N+1)^2(4N+1) - (4N+1)^2 = 4N^2(4N+1) = 4N^2M^2.$$

We note that the symmetric family (14.54) can also be deduced from (14.55) by adding $(4N + 2)^2$ and subtracting $(2N + 1)^2$ on both sides of (14.55), to get

$$\sum_{k=0}^{2N} k^2 + (2N+1)^2(4N+4) = \sum_{k=1}^{2N+1} (k + 2N + 1)^2. \tag{14.57}$$

Sums of squares with alternating signs.

The following identities, of a different type, are also of interest. They involve squares of consecutive integers with alternating signs.

$$-5^2 + 4^2 = -3^2$$

$$2^2 - 3^2 + 4^2 = -5^2 + 6^2$$

$$-11^2 + 10^2 - 9^2 + 8^2 = -7^2 + 6^2 - 5^2$$

$$4^2 - 5^2 + 6^2 - 7^2 + 8^2 = -9^2 + 10^2 - 11^2 + 12^2$$

$$-17^2 + 16^2 - 15^2 + 14^2 - 13^2 + 12^2 = -11^2 + 10^2 - 9^2 + 8^2 - 7^2$$

$$6^2 - 7^2 + 8^2 - 9^2 + 10^2 - 11^2 + 12^2 = -13^2 + 14^2 - 15^2 + 16^2 - 17^2 + 18^2.$$

The second, fourth, and sixth identities in this list are the special cases $n = 2, 4$, and 6 of the following family, in which n is even:

$$\sum_{k=0}^{n} (-1)^k (k+n)^2 = \sum_{k=1}^{n} (-1)^k (k+2n)^2. \tag{14.58}$$

The first, third, and fifth identities in the above list are the special members $n = 2, 4$, and 6 of the following family, in which n is even:

$$\sum_{k=0}^{n-1} (-1)^k (k+2n)^2 = \sum_{k=1}^{n-1} (-1)^k (k-2n)^2. \tag{14.59}$$

Both are special cases of a more general family valid for n even and arbitrary m:

$$\sum_{k=0}^{n-1}(-1)^k(k+m)^2+(2n)^2=-\sum_{k=0}^{n-1}(-1)^k(k+1+3n-m)^2. \tag{14.60}$$

Note that this family has a symmetry property with respect to the transition term $(2n)^2$. To illustrate with numerical examples, in (14.60) take $n=4$, and let $m=1$, 2, and $1/3$ to obtain:

$$1^2-2^2+3^2-4^2+8^2=-12^2+13^2-14^2+15^2$$
$$2^2-3^2+4^2-5^2+8^2=-11^2+12^2-13^2+14^2$$
$$1^2-4^2+7^2-10^2+24^2=-38^2+41^2-44^2+47^2.$$

The choice $m=1/3$ leads to sums with terms in arithmetic progression. This will always happen when m is rational.

To prove (14.60), write the sum on the left as the sum of $n/2$ differences:

$$m^2-(m+1)^2+(m+2)^2-(m+3)^2+\cdots+(m+n-2)^2-(m+n-1)^2.$$

Now factor each difference of squares, and this becomes

$$-(2m+1)-(2m+5)-\cdots-(2m+2n-3)=-mn-(1+5+\cdots+2n-3),$$

because the number of terms is $n/2$.

A similar treatment of the sum on the right transforms it to

$$\frac{n}{2}(6n-2m)+(3+7+\cdots+2n-1)=3n^2-mn+(3+7+\cdots+2n-1).$$

Therefore, the difference of the sum on the right of (14.60) and that on the left is

$$3n^2+(1+3+5+7+\cdots+2n-1)=4n^2,$$

which proves (14.60).

We also have identities with a weight factor r^2 in the transition term. For example, for even n and arbitrary r we have the symmetric family

$$\sum_{k=1}^{n-1}(-1)^k(2r^2n+k)^2+(\frac{2r^2n}{r})^2=\sum_{k=1}^{n-1}(-1)^k(2r^2n-k)^2. \tag{14.61}$$

For examples of (14.61), take $n=4$, and let $r=1,2$ and 3 to obtain

$$-11^2+10^2-9^2+(\frac{8}{1})^2=-7^2+6^2-5^2$$
$$-35^2+34^2-33^2+(\frac{32}{2})^2=-31^2+30^2-29^2$$
$$-75^2+74^2-73^2+(\frac{72}{3})^2=-71^2+70^2-69^2.$$

To prove (14.61) we write it in the equivalent form

$$(2n)^2r^2 = \sum_{k=1}^{n-1}(-1)^{k-1}[(k+2r^2n)^2 - (k-2r^2n)^2],$$

note that $(k+2r^2n)^2 - (k-2r^2n)^2 = 8r^2nk$, and then invoke the formula

$$\sum_{k=1}^{n-1}(-1)^{k-1}k = \frac{n}{2},$$

valid for even n.

A similar argument gives the following symmetric family of identities for even n and arbitrary r, with a new parameter $a = n(r^2-1)/2$:

$$\sum_{k=1}^{n-1}(-1)^k(k+a+2n)^2 + (2n)^2r^2 = \sum_{k=1}^{n-1}(-1)^k(2n-a-k)^2. \tag{14.62}$$

When $r = 1$, (14.62) becomes (14.59). The case $n = 6$ and $r = 1$ in (14.62) is displayed above as the case $n = 6$ in (14.59).

In (14.62), take $n = 6$, and let $r = 3$ and 4 to obtain

$$-41^2+40^2-39^2+38^2-37^2+12^2\cdot 3^2 = -(-13)^2+(-14)^2-(-15)^2+(-16)^2-(-17)^2$$

$$-62^2+61^2-60^2+59^2-58^2+12^2\cdot 4^2 = -(-34)^2+(-35)^2-(-36)^2+(-37)^2-(-38)^2.$$

In these two examples the symmetry is displayed by the fact that 12 is midway between 37 and (-13) in the first case, and midway between 58 and (-34) in the second case.

The foregoing families illustrate that a rich variety of identities can be easily obtained. For related work, see L. E. Dickson [37; pp. 318-323].

NOTES ON CHAPTER 14

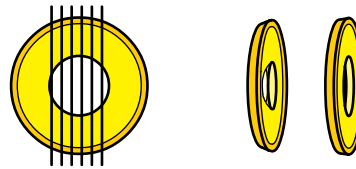
Most of the material in Sections 14.1 through 14.6 appeared in [11]. The identities in Section 14.7 appear in [27].

Chapter 15

APPENDIX

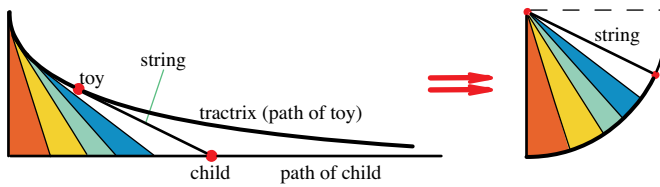
These problems can be easily solved by the methods developed in this chapter. The reader may wish to try solving them before reading the chapter.

1. The left part of the next figure shows a mid-section of a sphere with a concentric spherical cavity. The cavity is surrounded by two concentric spherical shells of arbitrary thicknesses, each having its own uniform density. Imagine the inner shell made of silver and the outer shell of gold. The sphere is sliced by parallel cuts into disks having the same width, two of which are shown on the right.



Prove that each punctured disk contains the same mass of gold and the same mass of silver, hence the same total mass of metal.

2. A tractrix is the curve traced by a toy being pulled on a taut string by a child walking along a fixed straight line, as shown below.



(a) *Explain why the diagram provides a proof without words that the area of the region under the tractrix is equal to that of a quarter of a circular disk whose radius is the length of the string.*

(b) *Find the area of the region under a curve traced by a knot in the middle of the string.*

CONTENTS

15

This appendix describes alternative approaches of some topics treated in earlier chapters, and points the way to further applications. For example, it relates the exponential curve and the tractrix, it introduces a geometric approach to tomography, and it discusses isoperimetric properties of circumgons. It also contains a proof of Mamikon's Theorem based on differential geometry.

15.1 ALTERNATIVE TREATMENT OF PARABOLIC SEGMENT

Figure 15.1 shows the parabola $y = x^2$, together with another parabola, $y = (2x)^2$, exactly half as wide as the first. Both are enclosed in a rectangle of base x and altitude x^2 whose area is x^3. The two parabolas divide the rectangle into three regions, and our strategy is to show that all three have equal areas. Then each has area one-third that of the circumscribing rectangle, as required.

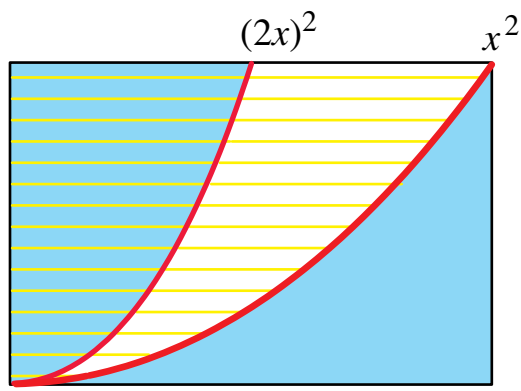


Figure 15.1: Parabola $y = (2x)^2$ formed by bisecting the horizontal segments of $y = x^2$.

The two leftmost regions formed by the bisecting parabola obviously have equal areas, so to complete the proof we need only show that the region above the bisecting parabola has the same area as the parabolic segment below the original parabola. To do this, look at Figure 15.2. The four right triangles in this figure have equal areas (they have the same altitude and equal bases). Therefore the problem reduces to showing that the two shaded regions have equal areas.

The shaded portion under the parabola $y = x^2$ is the tangent sweep obtained by drawing all the tangent lines to the parabola and cutting them off at the x axis. Next we show that the other shaded portion is its tangent cluster, with each tangent segment translated so that its point of intersection with the x axis is brought to a common point, the origin.

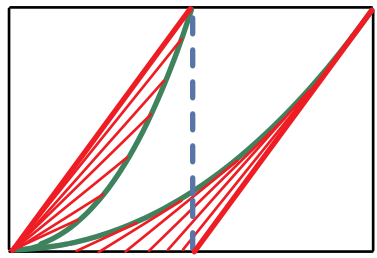


Figure 15.2: A tangent sweep and and its tangent cluster.

To see this, note that at a typical point (t, t^2) on the lower parabola, the tangent intersects the x axis at $t/2$. Therefore, if the tangent segment from $(t/2, 0)$ to (t, t^2) is translated left by $t/2$, the translated segment joins the origin and the point $(t/2, t^2)$ on the curve $y = (2x)^2$. So the tangent cluster of the tangent sweep is the shaded region above the curve $y = (2x)^2$, and by Mamikon's Theorem the two shaded regions have equal areas, as required. This completes the alternative proof that the area of a parabolic segment below the parabola is exactly one-third that of its circumscribing rectangle.

15.2 ALTERNATIVE TREATMENT FOR HIGHER POWERS

The argument used for a parabolic segment extends to generalized parabolic segments, in which x^2 is replaced by higher powers. Figure 15.3a shows the graphs of $y = x^3$ and $y =(3x)^3$, which divide the rectangle of area x^4 into three regions. The curve $y =(3x)^3$ trisects each horizontal segment in the figure, hence the area of the region above the cubic is half that of the region between the two cubic curves, as shown in the diagram.

Now we show that the area above the trisecting cubic is equal to the area below the other one, so each area is one-fourth that of the circumscribing rectangle.

To do this refer to Figure 15.3b, which shows the same two regions with a right triangle cut off each. Here we use the fact that the subtangent to the cubic is one-third the length of the base. So the two right triangles are congruent and have equal areas, and now we have to show that the two shaded regions have equal areas.

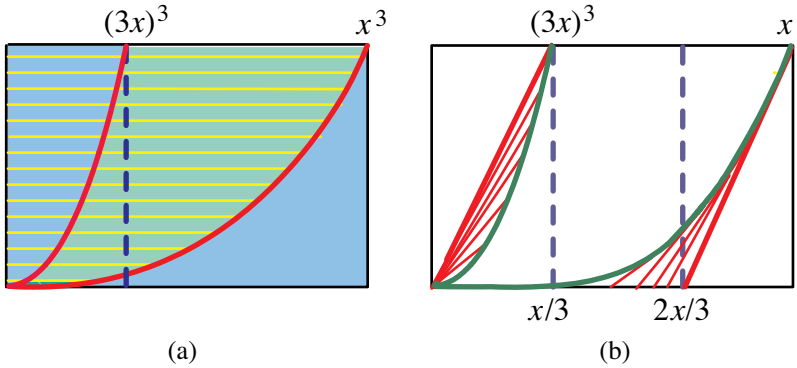


Figure 15.3: (a) Cubic curve $y = (3x)^3$. (b) The tangent sweep and the tangent cluster have equal areas.

One shaded region is the tangent sweep and the other is its tangent cluster, so they have equal areas. Since each of them has area half that of the region between the two cubics, the area of the cubic segment below the curve $y = x^3$ is one-fourth that of the rectangle, or $x^4/4$.

The argument extends to higher integer powers, a property not shared by the Archimedes treatment of the parabolic segment. For the curve $y = x^n$ we use the fact that the subtangent at x has length x/n and argue in a similar manner.

15.3 CALCULUS TREATMENT OF THE TRACTRIX

Mamikon's method of treating the tractrix problem is much simpler than the classical calculus treatment presented in this section.

Figure 15.4 shows a point (x, y) on a tractrix generated as the locus of a point moving in such a way that the tangent segment from (x, y) to the x axis has constant length k. The slope y' of the tangent line is $-\tan\theta$, where θ is the angle of inclination of the tangent segment with the x axis shown in Figure 15.4. The right

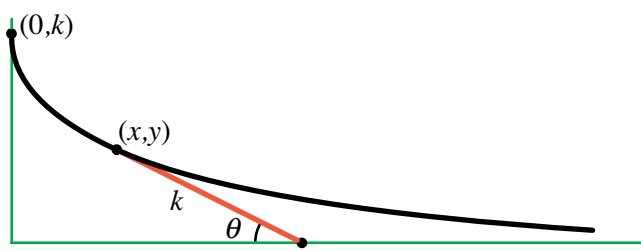


Figure 15.4: Determining the cartesian equation for the tractrix.

triangle with base angle θ and hypotenuse k has altitude y and base $\sqrt{k^2 - y^2}$.

Therefore $\tan\theta = y/\sqrt{k^2-y^2}$, hence y satisfies the differential equation

$$y' = -\frac{y}{\sqrt{k^2-y^2}}.$$

Writing $y' = dy/dx$, we can express this as a differential equation for x in terms of y :

$$\frac{dx}{dy} = -\frac{\sqrt{k^2-y^2}}{y}.$$

Integrating this with the help of the substitution $y = k\cos t$, and using the fact that $x = 0$ when $y = k$, we find

$$x(y) = k\log\frac{k+\sqrt{k^2-y^2}}{y} - \sqrt{k^2-y^2}. \tag{15.1}$$

This cartesian equation is satisfied by any point (x, y) on the tractrix, with $x = x(y)$. An alternative derivation of (15.1) is given in Chapter 11, Section 11.4.

The area of the region between the tractrix and the x axis is the integral $\int_0^k x(y)\,dy$, obtained by integrating the function in (15.1) with respect to y, from $y = 0$ to $y = k$. The indefinite integral of $x(y)$ can be determined by standard integration techniques. The result, which can be verified by differentiation, states that

$$\int (k\log\frac{k+\sqrt{k^2-y^2}}{y} - \sqrt{k^2-y^2}\,)\,dy = A(y),$$

where

$$A(y) = ky\log\frac{k+\sqrt{k^2-y^2}}{y} - \frac{1}{2}y\sqrt{k^2-y^2} + \frac{1}{2}k^2\arcsin\frac{y}{k}. \tag{15.2}$$

Now $A(0) = 0$, and $A(k) = \int_0^k x(y)dy$. Using (15.2) to calculate $A(k)$, we find that only the arcsine term survives and we get $A(k) = \pi k^2/4$. This is the area of a quarter of a circular disk of radius k, as predicted by Mamikon's Theorem.

15.4 GEOMETRIC DERIVATION OF THE INDEFINITE INTEGRAL

This section uses Figure 15.5 to derive (15.2) geometrically. In Figure 15.5a, the shaded region between the tractrix and the interval $[0, x]$, consists of two parts, a rectangle of area $yx(y)$ and a curved region above it of area $\int_y^k x(t)\,dt$. Let T denote the area of the adjacent right triangle with altitude y and hypotenuse of length k, tangent to the tractrix. As the tangent segment of length k moves along the tractrix from $(0, k)$ to (x, y) its tangent sweep, which consists of the shaded region together with the triangle, has area $\int_y^k x(t)\,dt + yx(y) + T$. By Mamikon's theorem this is equal to the area of the corresponding tangent cluster, a circular sector of radius k subtending angle θ shown in Figure 15.5b.

The cluster has area $\frac{1}{2}k^2\theta$. Therefore

$$\int_y^k x(t)\,dt + yx(y) + T = \frac{1}{2}k^2\theta,$$

which we can write as

$$\int_y^k x(t)\,dt + yx(y) = \frac{1}{2}k^2\theta - T. \tag{15.3}$$

The geometric meaning of (15.3) is revealed in Figure 15.5. The left side is the area of the darker shaded region in Figure 15.5a, which is the ordinate set of the tractrix over the interval $[0, x]$. The right side represents the area of the darker shaded region in Figure 15.5b: the area of the circular sector of radius k subtending angle θ, minus the area T of the right triangle.

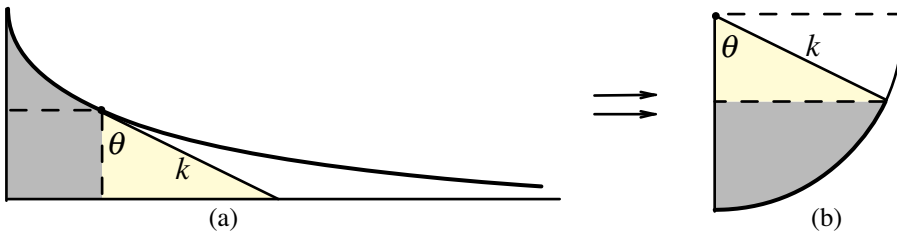


Figure 15.5: Geometric interpretation of (15.3).

We use this to derive the formula for $A(y)$ in (15.2). Write

$$A(y) = \int_0^y x(t)\,dt \;=\; (\int_0^k + \int_k^y)\,x(t)\,dt = A(k) - \int_y^k x(t)\,dt.$$

We know that $A(k) = \pi k^2/4$, and using (15.3) we find

$$A(y) = \frac{\pi k^2}{4} - \frac{1}{2}k^2\theta + T + yx(y). \tag{15.4}$$

To verify that this is the same as (15.2), we express θ and T in terms of k and y. Figure 15.5a gives us $\cos\theta = y/k$, or

$$\theta = \arccos\frac{y}{k} = \frac{\pi}{2} - \arcsin\frac{y}{k},$$

hence

$$-\frac{1}{2}k^2\theta = -\frac{\pi k^2}{4} + \frac{1}{2}k^2 \arcsin\frac{y}{k}.$$

For the triangular area T, we have

$$T = \frac{1}{2}yk\,\sin\theta \;= \frac{1}{2}y\,\sqrt{k^2 - y^2}.$$

Using these relations in (15.4) we find

$$A(y) = \frac{1}{2}k^2 \arcsin\frac{y}{k} + \frac{1}{2}y\sqrt{k^2 - y^2} + yx(y).$$

This has been obtained geometrically without integration or differentiation, and it is equivalent to (15.2), in view of (15.1).

15.5 SURPRISING RELATION BETWEEN EXPONENTIAL CURVES AND THE TRACTRIX

All subtangents from an exponential curve to the x axis have constant length, whereas all tangent segments from a tractrix to the x axis have constant length. This section shows that both curves are members of the same family in which a linear combination of the tangent and subtangent is constant.

Figure 15.6 shows an arbitrary curve together with a fixed base line, shown here as the x axis. At a point P of this curve a tangent segment of length t cuts off a subtangent of length s along the base line. As before, we can form the tangent cluster by translating each tangent segment of length t parallel to itself so the point of tangency is brought to a common point O, as in Figure 15.6 (right). Let C denote the other endpoint of the tangent.

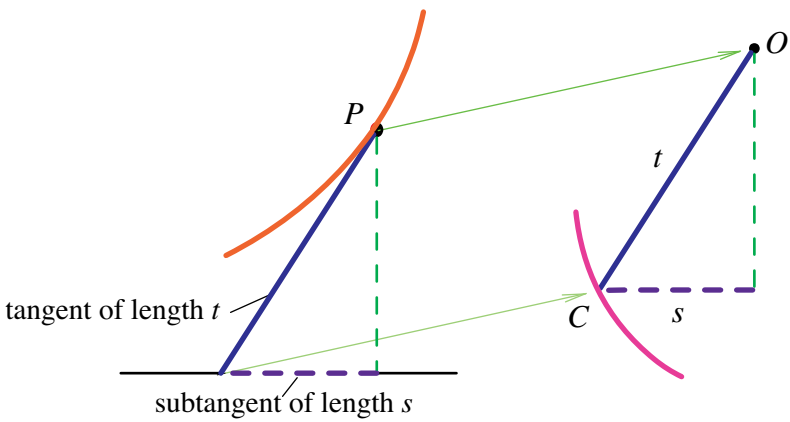


Figure 15.6: Tangent and subtangent of a general curve translated by the same amount.

As P moves along the curve, C traces the curve defining the tangent cluster. We can also translate the subtangent of length s. The subtangents will be parallel to the given base line. One endpoint of the translated subtangent is at C. When P moves along a tractrix, t is constant and C moves along a circle. When P moves along an exponential, s is constant and C moves along a vertical line.

Now suppose the original curve has the property that some linear combination of t and s has a constant value γ, say

$$\mu t + \nu s = \gamma$$

for some choice of nonnegative μ and ν, not both zero.

What is the path of C?

When $\nu = 0$, tangent t is constant and C lies on a circle. When $\mu = 0$, subtangent s is constant and C lies on a straight line. Now we show that, for general μ and ν, C always lies on a conic section. Let's see why.

If $\mu \neq 0$, divide by μ and rewrite the last equation as

$$t = e(d - s),$$

where $e = \nu/\mu$, and d is another constant. To show that C lies on a conic, refer to Figure 15.7. Use point O as a focus and take as directrix a line perpendicular to

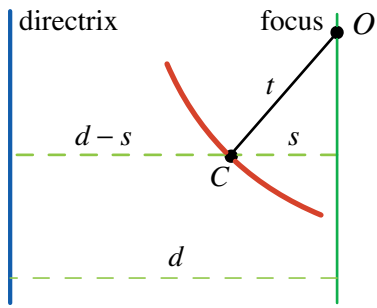


Figure 15.7: If $\mu t + \nu s$ is constant, C lies on a conic with eccentricity ν/μ.

the subtangents at a distance d from the focus and lying to the left of the vertical line through O. The quantity $(d - s)$ is the distance from the directrix to C, and t is the distance from the focus to C. The equation $t = e(d - s)$ states that the distance of C from the focus is e times its distance from the directrix, so C lies on a conic section with eccentricity e. The conic is an ellipse, parabola, or hyperbola according as $0 < e < 1$, $e = 1$, or $e > 1$. The limiting cases $e = 0$ (i.e., $\nu = 0$) and ∞ (i.e., $\mu = 0$) give a circle and vertical straight line, respectively.

An intermediate case occurs when the original point P lies on a pursuit curve in which a fox running on a line is pursued by a dog (not on the line) having the same speed as the fox. In this case it is easily shown that $t + s$ is constant, so $e = 1$ and the tangent cluster is a parabolic sector swept out by focal radii.

Thus, we have learned something interesting. The tractrix, the exponential, and the classical dog-fox-pursuit curve, which have been studied for hundreds of years, turn out to be special cases of a family of curves characterized by the equation $\mu t + \nu s =$ constant.

Cartesian equations for members of this family can be derived with the help of differential equations. However, they are not needed to determine the area of the tangent sweep of members of the family. By Mamikon's Theorem, the area of each tangent sweep is equal to that of the corresponding tangent cluster, which, in turn, is a sector of a conic section swept out by focal radii.

15.6 MORE GENERAL BICYCLE TRACKS

When bicycle tracks form a region like that in Figure 1.15, its area is that of a circular sector whose radius is the distance between the wheels of the bicycle, and whose central angle is the change in angle from the initial position to the final position.

Figure 15.8a shows a more general situation where the two tracks intersect to form several regions, labeled + and − to indicate that the rear wheel has crossed the path of the front wheel. The tangent sweep consists of the regions between the tracks, together with portions outside the tracks (labeled ± as in Figure 15.8b) swept twice, once in the positive (counterclockwise) direction, and again in the negative (clockwise) direction. The central angle of the tangent cluster gets positive contributions from regions marked +, and negative contributions from those marked − (Figure 15.8c). Contributions from common portions (marked ±) cancel, and again the change in angle depends only on the initial and final positions. Thus, the method of sweeping tangents gives the sum of the areas of regions marked +, minus the sum of the areas of regions marked − (Figure 15.8a).

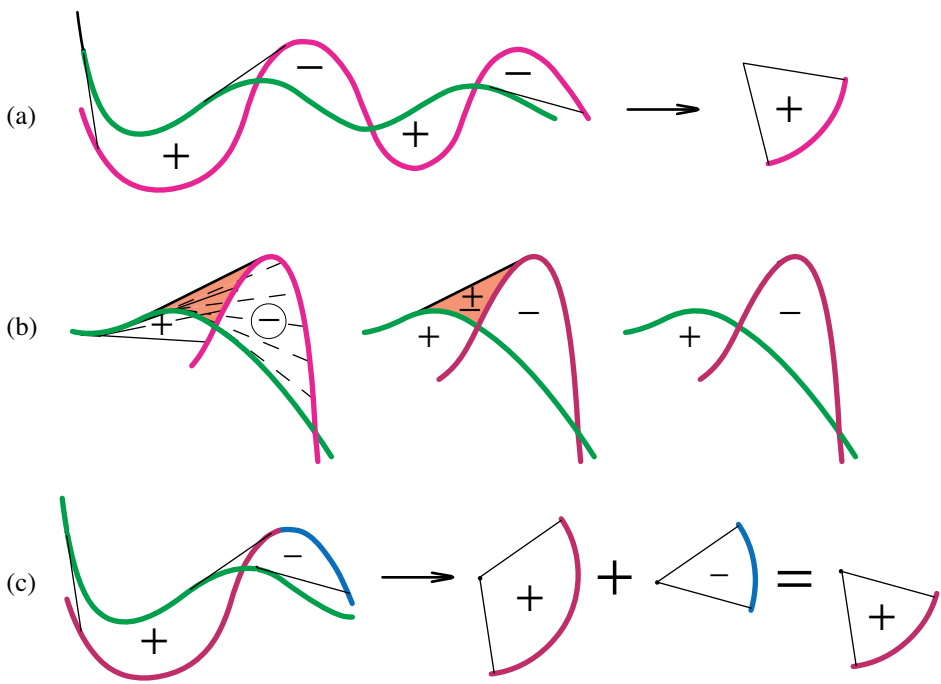


Figure 15.8: Intersecting tracks of bicycle wheels.

15.7 VARIATIONS ON THE TRACTRIX

Return to the tractrix in Figure 1.18, the trajectory of a toy on a taut string being pulled by a child walking along the x axis. Imagine a knot on the string and ask for its trajectory. If the knot is at the toy, the trajectory is the tractrix, and if it is at the child's hand, it is the linear path of the child.

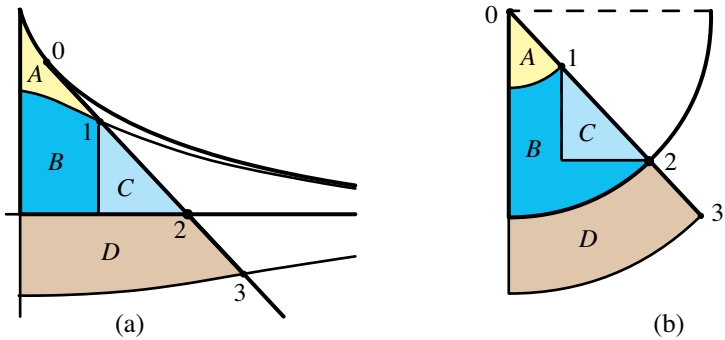


Figure 15.9: Trajectory of a knot on a tangent to a tractrix (a), and the corresponding tangent cluster (b). Regions labeled with the same letter in (a) and (b) have equal areas.

Other possible trajectories are shown in Figure 15.9a, including a case in which the knot is on an extended string below the x axis. In Figure 15.9a, point 0 is on the tractrix, point 2 is on the x axis, point 1 is between 0 and 2, and point 3 is collinear with 0, 1, 2, but below the x axis. A cartesian equation for each trajectory can be obtained from that of the tractrix. Areas of regions between trajectories can be determined without knowing the equations by consulting the corresponding tangent clusters in Figure 15.9b. For example, the area of the shaded region between the x axis and the trajectory through point 3 in Figure 15.9a is equal to the difference in areas of the corresponding sectors in Figure 15.9b.

The knot can also be placed beyond the toy as indicated in Figure 15.10. Its trajectory starts on the y axis, makes an arc to the left of the y axis, then returns to a lower point on the y axis and continues to the right as indicated.

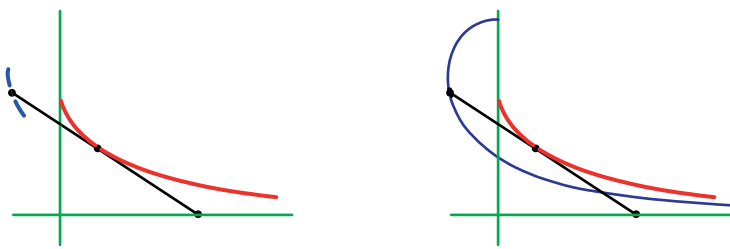


Figure 15.10: Knot beyond the toy and its trajectory.

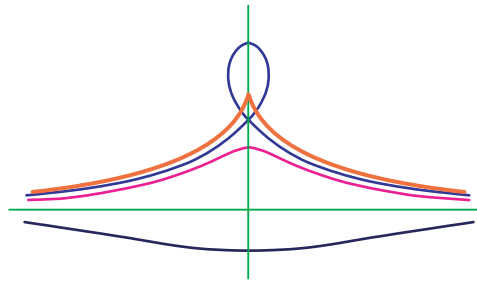


Figure 15.11: Symmetric tractrix and knot trajectories.

If the child is allowed to move along the negative x axis as well, the tractrix and the trajectories of the knots are symmetric about the y axis, as shown in Figure 15.11. When the knot is beyond the toy, the trajectory has a loop. The tractrix itself has a cusp on the y axis as indicated in Figure 15.11.

15.8 GEOMETRICAL APPROACH TO TOMOGRAPHY

Tomography reconstructs higher-dimensional density distribution of a solid body, such as a brain, from lower-dimensional projections, obtained by body-penetrating radiation such as X-rays. Modern tomography methodology relies on the numerical solution of millions of equations with millions of unknowns arising from complex mathematical analysis. By contrast, this section introduces a possible geometrical approach to tomography.

Spherical shells.

We begin with the simple case of spherically symmetric distributions. A spherical shell is the region between two concentric solid spheres. Its cross section by a plane that intersects the inner and outer spheres is an annular ring whose inner and outer radii depend on the cutting plane. In Section 1.3 we proved:

The area of an annular ring cut by a plane that intersects both spheres of a spherical shell is a constant independent of the position and inclination of the cutting plane.

This, in turn, was used in Chapter 5 to prove Theorem 5.4, which is restated here, and illustrated in Figure 15.12.

Theorem 5.4. *A slice of a spherical shell between two horizontal planes that cut both spheres has volume equal to the corresponding slice of a cylindrical shell cut by the same planes.*

Thus, for given radii, the volume of a slice of a spherical shell between two parallel planes that cut both spheres is proportional to the distance between the parallel cutting planes.

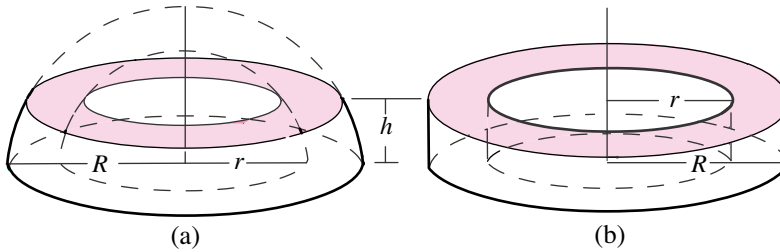


Figure 15.12: A slice of a spherical shell between two horizontal planes that cut both spheres has the same volume as the cylindrical shell cut by the planes.

Spherical distribution with uniform density.

Next, consider a uniform solid sphere whose density (mass per unit volume) is constant. We construct a corresponding projected mass density as follows. Slice the sphere by a finite number of equally spaced vertical parallel planes, as suggested by the vertical lines in Figure 15.13a, and draw a horizontal axis perpendicular to the cutting planes. We call this a projection axis, with its origin 0 on the vertical plane through the center of the sphere, and unit of length equal to the spacing between the cutting planes.

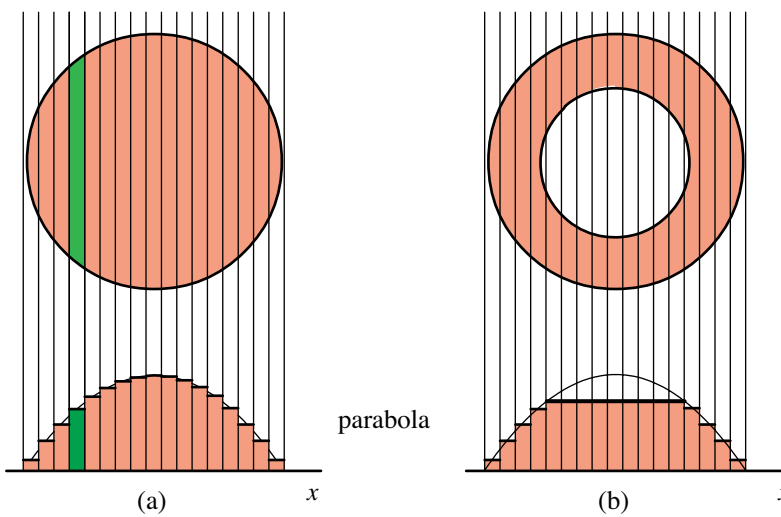


Figure 15.13: (a) Equally spaced slices of a uniform sphere. The histogram shows the projected mass density (mass/unit interval). (b) Uniform sphere with a cavity.

Form a histogram by constructing a step function whose height above each subinterval of unit length is the mass of the slice. Thus, the area of each rectangle represents the mass of the slice above its base. The histogram is symmetric because of spherical symmetry.

The cavity principle.

Start with a uniform spherical shell with outer radius r and a concentric spherical cavity of radius a, illustrated in Figure 15.13b. By Theorem 5.4, those slices in Figure 15.13b piercing the cavity have equal volume and equal mass (because the density is constant). Consequently, the portion of the histogram due to the cavity is constant, which means that the projected mass density is constant on the interval $[-a, a]$. We refer to this property as the *uniform cavity principle*:

For a given uniform spherical shell, the projected mass density due to the cavity is constant.

A simple but important consequence is illustrated in Figure 15.13b. The graph in Figure 15.13a represents the projected mass density of a uniform sphere of radius r with no cavity. Figure 15.13b shows the corresponding graph after a sphere of radius a is removed to create a cavity. The graph is constant over $[-a, a]$, and has the same shape as that in Figure 15.13a outside this interval. Therefore, the area of the region in Figure 15.13b above the horizontal line and below the graph of the full histogram is equal to the mass inside the sphere of radius a in Figure 15.13a from which the cavity was created. The uniform cavity principle and its consequence as described in Figure 15.13b can be extended to a spherically symmetric mass distribution, uniform or nonuniform, as follows.

General cavity principle. *For a sphere with spherically symmetric mass distribution and corresponding projected mass density histogram we have*:
(a) *The contribution to the histogram due to a cavity of radius a is a constant depending on this radius.*
(b) *The mass inside a concentric sphere of radius a is equal to the area of the portion of the full histogram above the constant value in* (a).

To establish this principle for any spherically symmetric distribution, we start with a sphere of radius a and regard it as a fixed cavity, around which we build a spherical shell by adjoining successive uniform concentric layers of different densities, like layers of an onion. In Figure 15.14a the radius of the outer sphere is very nearly the same as the radius a of the cavity, and we can think of this as a shell surrounding the cavity with a thin layer of uniform material. The corresponding histogram is shown as trapezoidal over the interval $[-a, a]$.

In Figure 15.14b, a second layer of uniform material has been added to the outer sphere in Figure 15.14a. The density in the new layer is also constant, but it need not be the same as that in the first layer. Again, the resulting histogram is horizontal above $[-a, a]$, but its constant height is greater by the constant amount of mass above coming from the second layer. Outside the interval $[-a, a]$, the histogram is correspondingly changed, but symmetry is preserved because the new distribution resulting from the two new layers is spherically symmetric.

In Figure 15.14c a third layer of uniform material has been added to the outer sphere in Figure 15.14b, and again the constant height of the histogram over the interval $[-a, a]$ is greater by the constant amount of mass coming from the third layer. As we continue this process, the resulting shells always have spherical symmetry, and the histogram above $[-a, a]$ is shown as a horizontal trapezoid at each

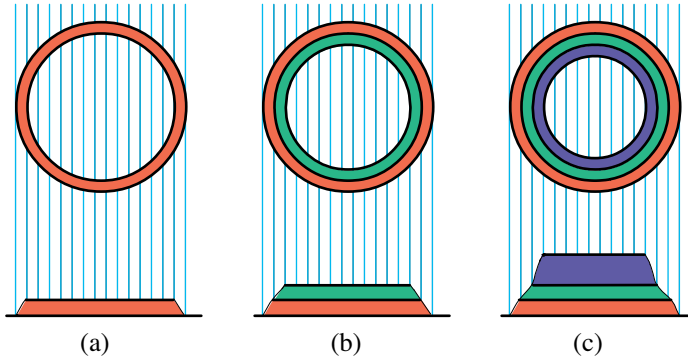


Figure 15.14: Spherically symmetric shells constructed with concentric uniform layers of different densities around a fixed cavity, together with their projected mass density trapezoidal histograms.

stage. Note that the area of each trapezoidal horizontal layer of the histogram above $[-a, a]$ is equal to the mass of the corresponding concentric layer projected onto the same interval. This process describes one way of building a spherically symmetric shell of nonuniform density around a cavity.

On the other hand, an arbitrary sphere having spherically symmetric nonuniform mass density can be visualized as being made up of layers of thin concentric uniform shells, each with its own constant density. The corresponding projected mass density histogram is shown in Figure 15.15a as a symmetric smooth curve; the area of its ordinate set represents the total mass in the solid sphere.

Now take a sphere, and remove a smaller concentric sphere to form a cavity of radius a, say. The corresponding projected mass density is constant over the interval $[-a, a]$, as shown by the example in Figure 15.15b, and otherwise has the same shape as the projected mass density in Figure 15.15a. Therefore, the area of

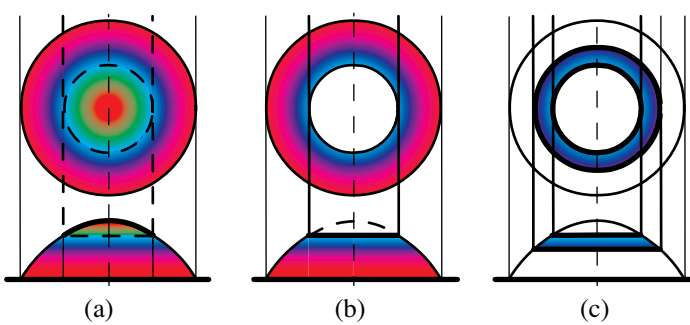


Figure 15.15: (a) Projected mass density of a spherically symmetric distribution. (b) The mass inside a sphere of radius a is equal to the area of the unshaded portion above $[-a, a]$. (c) Illustrating Theorem 15.1.

the portion cut off by this constant height, shown unshaded region in Figure 15.15b, is equal to the mass that was inside the sphere of radius a in Figure 15.15a before the cavity was created. This proves the general cavity principle stated above. As a consequence of this principle we can find the mass between any two spheres, as illustrated in Figure 15.15c.

Theorem 15.1. *The mass contained in a spherical shell is equal to the area of the corresponding horizontal slice of the one-dimensional density distribution.*

Let $\Phi(r)$ denote the three-dimensional mass density at distance r from the center of the shell, and let $f(x)$ denote the corresponding one-dimensional projected mass density function. Theorem 15.1 leads to a relation between the density function Φ and the derivative of f:

$$\Phi(r) = \frac{|f'(r)|}{2\pi r}. \tag{15.5}$$

To verify (15.5), note that a thin shell of radius r and thickness Δr has mass ΔM very nearly equal to $4\pi r^2 \Phi(r)\Delta r$. According to Theorem 15.1, ΔM is also equal to the area of the corresponding horizontal slice of the one-dimensional distribution, which is $2r|\Delta f|$, where $|\Delta f|$ is the altitude of the slice of base $2r$. Equate the two expressions for ΔM and let $\Delta r \to 0$ to obtain (15.5).

Figure 15.16 shows examples of f on $[-a, a]$, with $f(x) = 0$ if $|x| > a$. Figure 15.17 shows the general shape of the corresponding density Φ over the interval $[0, a]$. In Figure 15.17a, $f'(a)$ does not exist but the density Φ can be thought of as a Dirac delta function that is infinite at a and 0 elsewhere.

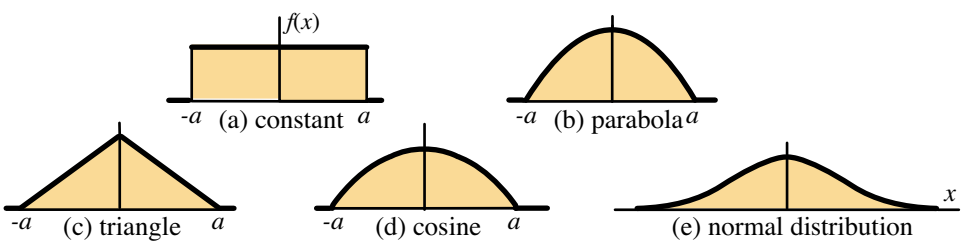


Figure 15.16: Examples of one-dimensional density distributions f.

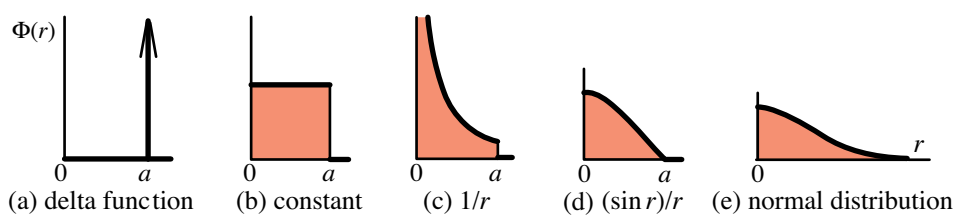


Figure 15.17: Corresponding three-dimensional density distribution Φ.

The foregoing geometric method can be further extended from spherically symmetric distributions to arbitrary asymmetric distributions. One can reconstruct an arbitrary three-dimensional mass density distribution from a knowledge of its one-dimensional projections in all possible directions, and thereby solve geometrically the central problem of classical tomography. However, this topic is beyond the scope of this book, and is not discussed here.

It is worth mentioning that the method for spherically symmetric distributions has already been applied to astrophysics. In [57] it was used to discover that the famous star cluster Pleiades contains a central cavity with respect to so-called flare stars, which sheds light on the evolution of young stars in stellar aggregates.

15.9 OPTIMAL CIRCUMGONS

Isoperimetric problems.

In the plane, traditional isoperimetric problems compare different plane regions having equal perimeters and ask for the region of maximal area. The analogy in 3-space is to compare different solids having equal surface area and ask for the solid of maximal volume. It is known that among all plane regions with a given perimeter, the circle encloses the largest area. This follows from the isoperimetric inequality $P^2 \geq 4\pi A$, which relates the area A and perimeter P of a planar region bounded by a simple closed curve. Equality holds only for the circle.

A well known theorem [61; Theorem 12.5a] states:

A polygon inscribed in a circle has larger area than any other polygon with sides of the same lengths (*therefore of the same perimeter*) *in the same order.*

The following is a dual to this statement that to our knowledge has never been previously formulated:

Theorem 15.2. *A polygon circumscribing a circle has a larger area than any other polygon with the same perimeter and the same interior angles in a given order.*

This section outlines a method of proof that also reveals interesting isoperimetric properties of polygonal circumgons along the way. Our principal tool is the concept of contour ratio, introduced in Chapter 10, a number associated with a plane region having perimeter P and area A. It is defined as the quotient

$$\kappa = \frac{P^2}{4A}, \tag{15.6}$$

and has the pleasant feature that $\kappa = \pi$ for a circle and $\kappa = 4$ for a square. The isoperimetric inequality states that $\kappa \geq \pi$ for any plane region bounded by a simple closed curve. It also implies that among all regions with a given area, the circle has the smallest perimeter.

The ratio π/κ is called the isoperimetric quotient, terminology that Niven [61; p. 90] attributes to Pólya.

Similar contours have the same contour ratio because the scaling factor cancels in (15.6). For contours with given perimeter, those with larger areas have smaller

contour ratios. And for contours with given area, those with smaller perimeters have smaller contour ratios.

A detailed study of contour ratios is given in Chapter 10. Here we use only the simple fact that for a circumgon with inradius r and contour ratio κ, we have

$$P = 2\kappa r \text{ and } A = \kappa r^2. \tag{15.7}$$

These equations follow at once from (15.6) because $A = Pr/2$ for a circumgon. They generalize the classical formulas $P = 2\pi r$ and $A = \pi r^2$ for a circular disk. Thus, the contour ratio κ plays the same role for a circumgon that π plays for a circle; it is the ratio of the perimeter to the diameter of the incircle. For a given inradius, both P and A are proportional to κ, so they increase or decrease with κ.

Some isoperimetric problems deal with specific contours, such as polygons. An n-gon has n edges and n interior angles. For a given n we can vary the lengths of the edges or the measures of the angles and try to minimize the contour ratio. A classical result, attributed to Zenodorus [46; p. 207, (3)], states that among all n-gons of given perimeter, the regular polygon is greatest in area. Thus, for a triangle of given perimeter, the equilateral triangle has maximal area, and for a quadrilateral with given perimeter the square has maximal area. It is also well known that among triangles with one given angle the isosceles triangle has the smallest contour ratio.

We can vary a triangle and preserve its angles by moving each side parallel to itself. This produces a similar triangle that necessarily has the same contour ratio. In other words, if the angles of a triangle are fixed, and each side is moved parallel to itself, the contour ratio remains constant. Motivated by the example of a triangle, which is a circumgon, we ask: *What happens to the contour ratio if each side of a general circumgon is moved parallel to itself*?

First we prove a simple but powerful auxiliary result relating the contour ratio of a circumgon with that of an arbitrary contour. Start with a circumgon with perimeter P, area A, and contour ratio $\kappa = P^2/(4A)$. Now take a contour with perimeter P' and area A', and compare its contour ratio

$$\kappa' = \frac{(P')^2}{4A'}$$

with that of the circumgon. There are three possibilities: $\kappa' > \kappa$, $\kappa' = \kappa$, and $\kappa' < \kappa$. The following lemma expresses each in an equivalent form that involves the quantitites P, A, P', A' and the inradius r of the circumgon.

Comparison Lemma. *Let κ denote the contour ratio of a circumgon with perimeter P, area A, and inradius r, and let κ' denote the contour ratio of a contour with perimeter P' and area A'. Let $\otimes$ denote one of the three order relations $>$, $=$, or $<$. Then the relation*

$$\kappa' \otimes \kappa \tag{15.8}$$

is equivalent to the following relation involving the differences $P' - P$ and $A' - A$:

$$r^2(P' - P)^2 + 4Ar(P' - P) \otimes 4A(A' - A). \tag{15.9}$$

Proof. First, (15.8) can be written as $(P')^2 \otimes 4\kappa A'$, which is equivalent to

$$(P')^2 - 4\kappa A \otimes 4\kappa(A' - A). \tag{15.10}$$

Using (15.7) we have

$$(P')^2 = (P' - P + 2\kappa r)^2 = (P' - P)^2 + 4\kappa r(P' - P) + 4\kappa A,$$

or

$$(P')^2 - 4\kappa A = (P' - P)^2 + 4\kappa r(P' - P).$$

Use this in the left-hand side of (15.10), then multiply both sides by r^2, and again use $\kappa r^2 = A$ to get (15.9).

The Comparison Lemma is particularly adaptable for studying the behavior of the contour ratio when each edge of a circumgon is moved parallel to itself.

We introduce some convenient terminology. Recall that a polygonal circumgon is the union of a finite number of triangular building blocks, with each outer edge lying on a line tangent to the incircle. In what follows we assume the polygon is closed, which means that consecutive outer edges intersect, like those in Figures 4.1 and 4.2. The example in Figure 4.4 is not closed. We say that two consecutive outer edges form a convex corner if the interior angle of intersection is less than a straight angle, and a nonconvex corner if the angle of intersection is greater than a straight angle. In Figure 4.1, all corners are convex. In Figure 4.2a, four corners are convex and two are nonconvex, and in Figure 4.2b there are five convex corners and five nonconvex corners.

Our strategy involves starting with a given (closed) polygonal circumgon and forming a new closed polygon, which we call a parallel polygon, whose edges are along the same tangent lines or parallel to those of the given circumgon. This parallel polygon is constructed as follows: Imagine the line through an outer edge of the circumgon translated (parallel to itself). The two tangent lines through the outer edges adjacent to this segment are kept fixed. They intersect the translated line in a segment parallel to the outer edge. The new polygon, formed by replacing the outer edge of the circumgon by the translated segment, will have all its edges parallel to those of the circumgon. The length of the translated segment will depend on the angle of inclination of the two adjacent tangent lines and on the distance translated.

Now we can prove the following result for convex polygonal circumgons:

Theorem 15.3. *If one edge of a convex polygonal circumgon with contour ratio κ is translated to form a parallel polygon with contour ratio κ', then $\kappa' \geq \kappa$, with equality if and only if the circumgon is a triangle.*

Proof. As noted earlier, equality holds for triangles because parallel triangles are similar, hence have the same contour ratio. Therefore we consider convex polygonal circumgons with four or more sides. We form a parallel polygon by moving just one side of the circumgon parallel to itself, and show that we obtain strict inequality $\kappa' > \kappa$.

First we dispose of the simplest case, when the two edges adjacent to the moving edge are parallel to each other. Then the moving edge has constant length $a > 2r$, where r is the inradius. If the edge moves a distance w, the area A' and perimeter P' of the parallel polygon are given by

$$A' = A \pm aw, \;\; P' = P \pm \frac{a}{r}w,$$

with the plus or minus sign depending on the direction of translation. When these are used in (15.9) the Comparison Lemma tells us that $\kappa' \otimes \kappa$ is equivalent to $a^2w^2 \otimes 0$. Therefore the symbol $\otimes$ must be $>$, hence $\kappa' > \kappa$, which means that the contour ratio increases, regardless of the direction of translation.

If the two edges adjacent to the moving edge intersect, we consider two cases, depending on whether the length of the moving edge (a) decreases or (b) increases as it moves away from the incenter as suggested by edge MN in Figure 15.18.

Case (a). Start with an edge of the circumgon between two (convex) corners as shown in Figure 15.18, let a denote its length, and let a' denote the length of the translated parallel edge, which depends on w, the perpendicular distance between the parallel edges. In this case a' decreases as the new edge moves away from the incenter (and increases as it moves towards the incenter). If the new edge is outside the circumgon, as shown in Figure 15.18, w is restricted by the inequalities $0 \le w \le H$, where H is the altitude of the triangle with base a and opposite vertex at the point of intersection of the extended tangent lines adjacent to the edge of length a. In this case the parallel polygon has area

$$A' = A + \frac{a + a'}{2}w, \tag{15.11}$$

where $(a+a')w/2$ is the area of the trapezoid formed by the parallel edges of lengths a and a'. To determine how a' depends on w, use similar triangles to get

$$\frac{a'}{a} = \frac{H - w}{H} = 1 - \frac{w}{H},$$

so $a' = a - wa/H$, $a' + a = 2a - wa/H$, and (15.11) gives us

$$A' - A = aw - \frac{a}{2H}w^2. \tag{15.12}$$

Next, we show that the new perimeter P' is related to P by

$$P' - P = \frac{a}{r}\, w, \tag{15.13}$$

where r is the inradius. To see this, note that $P' = P'(w)$ is a linear function of w that has the value P when $w = 0$ so $P'(w) - P = cw$ for some constant c. To determine c, take $w = H$ to get $P'(H) - P = cH$. But when $w = H$ we have

$$P'(H) - P = e + f - a,$$

where e and f are the lengths (Figure 15.18) of the other two edges of the triangle with base a, altitude H, and area $T = aH/2$. The area T is also the sum of the

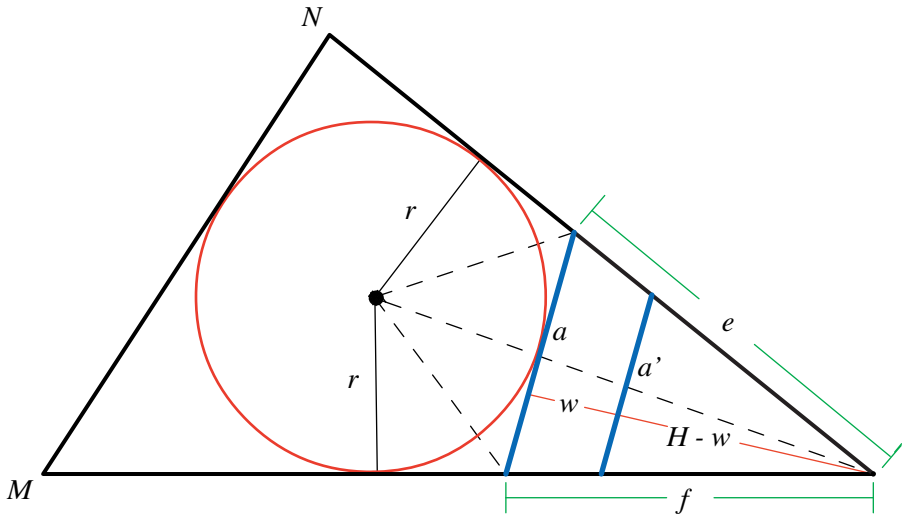


Figure 15.18: One edge between two-convex corners translated parallel to itself through a distance w to form a parallel polygon.

areas $er/2 + fr/2$ of the two triangles with one vertex at the incenter and opposite edges of lengths e and f, minus the area $ar/2$ of the overlapping triangle with one vertex at the incenter and opposite edge of length a. In other words, we have

$$aH/2 = (e + f - a)r/2 = cHr/2,$$

hence $c = a/r$ and we get (15.13).

If the parallel edge, of length a', is inside the circumgon, the same type of analysis gives

$$A' - A = -aw - \frac{a}{2H}w^2, \quad P' - P = -\frac{a}{r}w. \tag{15.14}$$

These are like (15.12) and (15.13) except for the sign change in the linear w term. Using them in the Comparison Lemma we find that $\kappa' \otimes \kappa$ is equivalent to

$$r^2(\pm\frac{a}{r}w)^2 + 4Ar(\pm\frac{a}{r}w) \otimes 4A(\pm aw - \frac{a}{2H}w^2),$$

where the plus sign is used if the translated edge of length a' is outside the circumgon, and the minus sign if it is inside. In either case, the linear terms in w cancel and the last relation simplifies to $a^2 \otimes -2\,\frac{aA}{H}$, or $aH/2 \otimes -A$, which becomes

$$T \otimes -A, \tag{15.15}$$

where $T = aH/2$ is the area of the triangle with base a and altitude H. Both T and A are positive, hence (15.15) is satisfied only when $\otimes$ is the symbol $>$, which means

$\kappa' > \kappa$. In other words, moving the new edge outward or inward will increase the contour ratio.

Case (b). Next, we move the edge labeled MN in Figure 15.18 parallel to itself and apply the same analysis. In this case, a' increases as the new edge moves away from the incenter (and decreases as it moves towards the incenter). Then, instead of (15.15), we find $T \otimes A$, where T is the area of the large triangle with base MN and altitude H perpendicular to MN. But in this case the large triangle contains the entire circumgon of area A, so $T > A$, and again we find $\kappa' > \kappa$.

In both cases (a) and (b) we used the Comparison Lemma to transform the relation $\kappa' \otimes \kappa$ between contour ratios to an equivalent relation $T \otimes \pm A$ between the area T of a triangle and the area A of the circumgon. The same method can be applied to a more general situation in which the edges of a convex polygonal circumgon undergo parallel translation independently through respective distances $w_1, \ldots, w_n$. The details are more complicated, but the essential ideas and final conclusion are the same as above: the contour ratio decreases. Consequently, we have

Theorem 15.4. *A convex polygonal circumgon has the minimal contour ratio among all parallel polygons obtained by moving independently one or more edges of the circumgon.*

If we move more than one edge of a polygonal circumgon, the resulting parallel polygon could become a similar scaled version of the original circumgon, in which case it has the same contour ratio, and its incircle is correspondingly scaled.

Nonconvex circumgons.

Theorem 15.4 does not apply to a nonconvex polygonal circumgon as we can easily verify by an example. Figure 15.19 shows a circumgon where two adjacent outer edges meet at an interior angle greater than a straight angle to form a nonconvex corner. As in the foregoing proof, we translate the line through the edge of length a parallel to itself through a distance w to produce a parallel polygon with a parallel edge of length a'. The new edge can be outside the circumgon, in which case $a' > a$, as in Figure 15.19, or inside the circumgon, in which case $a' < a$.

If A and P denote the area and perimeter of the circumgon, an argument similar to that given in Case (a) of Theorem 15.3 shows that area A' and perimeter P' of the parallel polygon satisfy the relations

$$A' - A = \pm aw + \frac{a}{2H}w^2 \tag{15.16}$$

and

$$P' - P = \pm\frac{a}{r}w, \tag{15.17}$$

where r is the inradius and H is the altitude of the triangle with base of length a. The plus sign is used if the translated edge is outside the circumgon, and the minus sign if it is inside. These are the same relations obtained in the proof of Theorem 15.3, except for a sign change in the term involving w^2. Consequently,

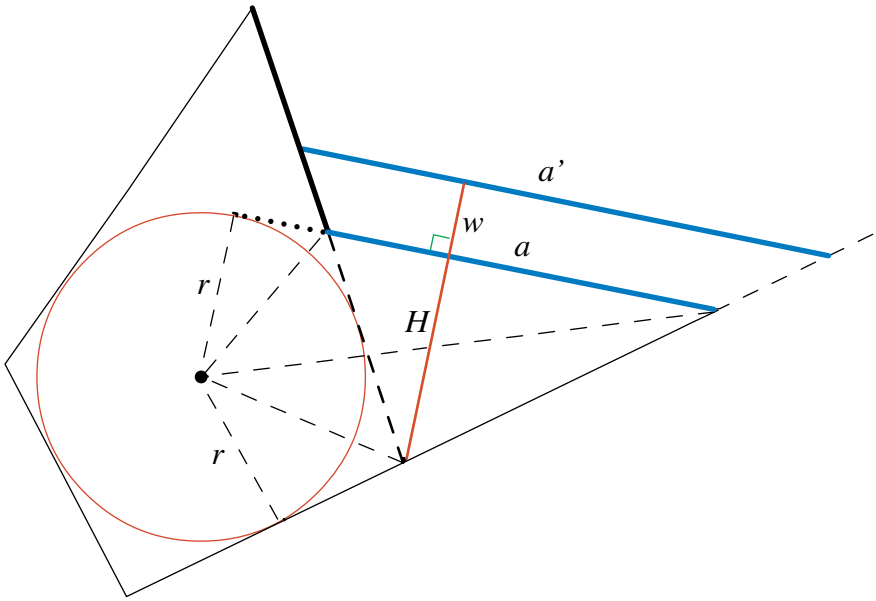


Figure 15.19: One edge at a nonconvex corner. Moving the line through this edge parallel to itself through a distance w forms a parallel polygon.

when (15.16) and (15.17) are used in the Comparison Lemma, we find the linear terms in w cancel as before, but instead of (15.15) we are left with the relation

$$T \otimes A, \tag{15.18}$$

where T is the area of the triangle of base a and altitude H. If the triangle is a proper subset of the circumgonal region, as happens in Figure 15.19, then (15.18) can be satisfied only when the symbol $\otimes$ is $<$. In other words, in this circumstance, the new polygon has contour ratio $\kappa' < \kappa$, which means that translating an edge at a nonconvex corner in either direction decreases the contour ratio. Intuitively, this is to be expected because the motion reduces the effect of nonconvexity.

If, in Figure 15.19, a different edge between two convex corners is moved, then it is easy to see that the contour ratio may increase or decrease, depending on the relation $T \otimes A$. Thus, for a nonconvex circumgon, moving an edge parallel to itself could produce a parallel polygon whose contour ratio is greater than, equal to, or less than that of the original circumgon. Summarizing, we have:

Theorem 15.5. *Assume one edge of a nonconvex polygonal circumgon with contour ratio κ is translated to form a parallel polygon with contour ratio κ'.*

(a) *If the translated edge is at a nonconvex corner, then $\kappa' < \kappa$.*

(b) *If the translated edge is between two convex corners, the relation $\kappa' \otimes \kappa$ is equivalent to $T \otimes A$, where T and A are the areas described above.*

In both cases (a) and (b), the relation $\kappa' \otimes \kappa$ between the contour ratios is

reduced to an equivalent relation $T \otimes A$ between areas, as in Case(b) of Theorem 15.3. The power of Theorem 15.5(b) lies in the fact that the complex behavior of the contour ratio in the nonconvex case is reduced to simply comparing the area of a triangle with that of the circumgon. This makes it possible to construct examples of different types in which the relation is $>, =$ or $<$. We have not explored what happens to the contour ratio if more than one edge of a nonconvex circumgon is translated.

Our original approach used multivariate calculus, treating the difference $\kappa' - \kappa$ as a function of the independent distances $w_1, \ldots, w_n$ through which the n edges were translated. By equating various partial derivatives to zero, we found that the circumgon gave a local extremum, but to determine whether the extremum is a maximum or minimum involved determining the algebraic sign of a quadratic form involving second-order derivatives. From our discussion, we see that the quadratic form is of fixed sign in the convex case, but need not be of fixed algebraic sign if the circumgon is nonconvex. By using the Comparison Lemma, as was done in proving Theorem 15.5, the argument is more elementary and more transparent than one using multivariate calculus.

Zenodorus revisited.

We conclude this section with an application to a result attributed to Zenodorus:

Among all n-gons of given perimeter, the regular n-gon has largest area.

We will prove the equivalent statement:

Among all n-gons, the regular n-gon has smallest contour ratio.

Suppose we have a convex polygonal circumgon. If it is equiangular, all interior angles are equal (have the same measure). Otherwise, at least one edge intersects its adjacent tangent lines to form unequal interior angles. We can change its direction, keeping it tangent to the incircle, until the two adjacent interior angles are equal. This gives a new circumgon with the same incircle that has a smaller area and hence a smaller contour ratio (because $A = \kappa r^2$). To see why the area is smaller, assume first that the adjacent tangent lines are not parallel. Then the edge in question is the base of a protruding triangle whose opposite vertex is the point of intersection of the two adjacent tangent lines. It is part of a circumgon with one less edge but having the same incircle, and we obtain the original circumgon by chopping it off. This triangle has the same perimeter, regardless of the direction of the chopping line (the perimeter is the sum of the lengths of the two fixed tangent segments from the incircle to the point of intersection), so if we remove the triangle of largest area, the leftover circumgon will have the smallest area.

Which direction should the chopping line have to remove the triangle of largest area? As already mentioned, for a triangle with one given angle and fixed perimeter, the isosceles triangle has the smallest contour ratio hence the largest area. A similar argument applies in two other situations, when the adjacent sides are parallel, and when they intersect at the other side of the circumgon, but we omit the details. In all cases, moving the edge to produce equal interior angles will decrease the area

and hence the contour ratio of the circumgon. This proves, by contradiction, the result of Zenodorus, because an equiangular circumgon is also a regular polygon.

15.10 PROOF OF MAMIKON'S THEOREM

This section uses differential geometry to prove Mamikon's Theorem. Start with a smooth space curve Γ described by a position vector $\boldsymbol{X}(s)$, where s, the arc-length function for the curve, varies over an interval, say $0 \le a \le s \le b$. The unit tangent vector to Γ is the derivative $d\boldsymbol{X}/ds$, which we denote by $\boldsymbol{T}(s)$. The derivative of the unit tangent is given by

$$\frac{d\boldsymbol{T}}{ds} = \kappa(s)\boldsymbol{N}(s),$$

where $\boldsymbol{N}(s)$ is the principal unit normal and $\kappa(s)$ is the curvature.

The curve Γ generates a surface S that can be represented by the vector parametric equation

$$\boldsymbol{y}(s,u) = \boldsymbol{X}(s) + u\boldsymbol{T}(s),$$

where u varies over an interval whose length can vary with s, say $0 \le u \le f(s)$. As the pair of parameters $(u,\ s)$ varies over the ordinate set of the function f over the interval $[a,\ b]$, the surface S is swept out by tangent segments extending from the initial curve Γ to another curve described by the position vector $\boldsymbol{y}(s,\ f(s))$.

Geometrically, S is a developable surface, that is, it can be rolled out flat on a plane without distortion. We refer to the surface S generated from Γ in this fashion as a tangent sweep.

The area of S is given by the double integral

$$a(S) = \int_a^b \int_0^{f(s)} || \frac{\partial \boldsymbol{y}}{\partial s} \times \frac{\partial \boldsymbol{y}}{\partial u} || \, du\, ds.$$

To calculate the integrand, we have

$$\frac{\partial \boldsymbol{y}}{\partial s} = \frac{\partial \boldsymbol{X}}{\partial s} + u\,\frac{d\boldsymbol{T}}{ds} = \boldsymbol{T}(s) + u\kappa(s)\boldsymbol{N}(s),$$

$$\frac{\partial \boldsymbol{y}}{\partial u} = \boldsymbol{T}(s), \qquad \frac{\partial \boldsymbol{y}}{\partial s} \times \frac{\partial \boldsymbol{y}}{\partial u} = u\kappa(s)\,\boldsymbol{N}(s) \times \boldsymbol{T}(s),$$

so

$$|| \frac{\partial \boldsymbol{y}}{\partial s} \times \frac{\partial \boldsymbol{y}}{\partial u} || = u\kappa(s),$$

because $||\boldsymbol{N}(s) \times \boldsymbol{T}(s)|| = 1$. The integral for the area becomes

$$a(S) = \int_a^b \Big(\int_0^{f(s)} u\, du \Big)\, \kappa(s)\, ds = \frac{1}{2} \int_a^b f^2(s)\kappa(s)\, ds.$$

Next, imagine the arc length s expressed as a function of the angle φ between the tangent vector $\boldsymbol{T}$ and a fixed tangent line, say the tangent line corresponding to $s = a$. When s is expressed in terms of φ, the function $f(s)$ becomes a function

of φ, and we write $f(s) = r(\varphi)$. On the surface S, φ is the angle between tangent geodesics, so the curvature κ is the rate of change of φ with respect to arc length, $\kappa = d\varphi/ds$. In the last integral we make a change of variable expressing s as a function of φ. Then $f^2(s) = r^2(\varphi)$, $\kappa(s)ds = d\varphi$, and the integral for $a(S)$ becomes

$$a(S) = \frac{1}{2} \int_{\varphi_1}^{\varphi_2} r^2(\varphi)\, d\varphi, \tag{15.19}$$

where φ_1 and φ_2 are the initial and final angles of inclination corresponding to $s = a$ and $s = b$, respectively. Formula (15.19) shows that the area $a(S)$ does not depend explicitly on the arc length of Γ at all; it depends only on the angles φ_1 and φ_2. In fact, $a(S)$ is equal to the area of a plane radial set with polar coordinates (r, φ) satisfying $0 \le r \le r(\varphi)$ and $\varphi_1 < \varphi \le \varphi_2$.

When (15.19) is reformulated in geometric terms, it yields Mamikon's Theorem in a form that has a strong intuitive flavor. If we translate each tangent segment of length $r(\varphi)$ parallel to itself so that each point of tangency is brought to a common vertex O, we obtain a portion of a conical surface that we call the tangent cluster of the curve Γ. Then (15.19) gives us:

Mamikon's Theorem. *The area of a tangent sweep of a curve is equal to the area of its corresponding tangent cluster.*

The proof for special curves of the type described above follows from (15.19). The tangent cluster of S lies on a conical surface that can be unrolled without distortion of area, and the unrolled tangent cluster becomes a plane region whose area in polar coordinates is given by (15.19). Since this is also the area of S, the tangent sweep and its tangent cluster have equal areas. The theorem is also true for more general surfaces that can be decomposed into a sum or difference of a finite number of special surfaces of the type in the foregoing discussion. This will take care of tangent sweeps generated by piecewise smooth curves. For example, polygonal curves are treated in Figure 1.10, and in [55]. It also takes care of curves with inflection points, where the sweeping tangent segments change their direction of rotation.

15.11 ARCHIMEDES' LAW OF THE LEVER

The famous *law of the lever*, discovered by Archimedes, states that:

Two weights are in equilibrium about a fulcrum if placed at distances inversely proportional to the weights.

In other words, if weights A and B are placed at respective distances a and b from the fulcrum, they will be in equilibrium if and only if

$$Aa = Bb. \tag{15.20}$$

Here we deduce (15.20) by considering weight distribution along a uniform horizontal rod. "Uniform" means that the weight of any portion of the rod is a constant times its length. There is no loss of generality in assuming the constant is 1, so

the weight of any portion can be equated to its length. Figure 15.20 illustrates the following property:

A uniform rod of finite length is in equilibrium when balanced on a fulcrum at its midpoint, which is its center of gravity.

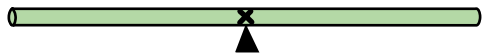


Figure 15.20: Uniform rod balanced at its center.

Imagine such a rod, balanced at its midpoint, and divided arbitrarily into two pieces of lengths A and B, as shown in Figure 15.21a.

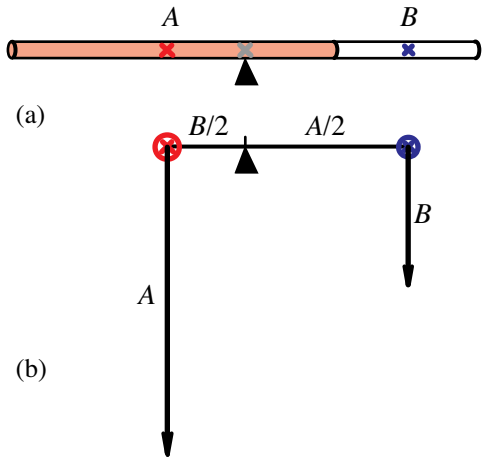


Figure 15.21: (a) Rod divided into pieces of lengths A and B. (b) Each piece is replaced by its weight placed at distances $B/2$ and $A/2$ from the fulcrum.

The concept of center of gravity tells us that equilibrium is maintained if a weight A is placed at the center of gravity of the left piece, and a weight B is placed at the center of gravity of the right piece. Now we show that the respective distances from the fulcrum are $a = B/2$ and $b = A/2$, as indicated in Figure 15.21b.

To do this, imagine a horizontal x axis with $x = 0$ at the left end of the rod. The right end is at $x = A + B$, and the center of gravity of the entire rod is at $x = (A + B)/2$, its midpoint. The left piece, of length A, has its center of gravity at $x = A/2$, which is at a distance $B/2$ to the left of the fulcrum. This tells us that if we remove the piece of length B, the new center of gravity shifts to the left by a distance $B/2$ from the fulcrum. Therefore $a = B/2$. If, instead, we remove the piece of length A, the new center of gravity shifts to the right by a distance $A/2$ from the fulcrum, hence $b = A/2$.

Now we note that the law of the lever in (15.20) is equivalent to the equation $A \cdot (B/2) = B \cdot (A/2)$. This simple demonstration uses the ideas of Archimedes, but the presentation is more transparent than that given in [44, p. 192].

NOTES ON CHAPTER 15

The alternative treatment for determining the area of the parabolic segment and of the region under the graph of a general power function was first published in [3]. The surprising relation between the exponential curve and the tractrix first appeared in [8]. The proof of Mamikon's Theorem given in Section 15.10 is essentially the same as that given in [59], [8], and [22]. The material in Sections 15.6 through 15.9 and that in Section 15.11 has not been previously published.

Bibliography

1. T. M. Apostol, *Calculus: vol. 1,* 2nd ed., John Wiley & Sons, New York, 1967.
2. ——, *Calculus: vol. 2,* 2nd ed., John Wiley & Sons, New York, 1969.
3. ——, A visual approach to calculus problems, *Engineering and Science*, vol. LXIII, no. 3, (2000), 22–31.
4. T. M. Apostol and M. A. Mnatsakanian, Surprising geometric properties of exponential functions, *Math Horizons*, (September 1998), 27–29.
5. ——, Cycloidal areas without calculus, *Math Horizons*, (September 1999), 12–18.
6. ——, Finding centroids the easy way, *Math Horizons*, (September 2000), 7–12.
7. ——, Sums of squares of distances, *Math Horizons*, (November 2001), 21–22.
8. ——, Subtangents—an aid to visual calculus, *Amer. Math. Monthly*, 109 (June/July 2002), 525–533.
9. ——, Generalized cyclogons, *Math Horizons,* (September 2002), 25–29.
10. ——, Tangents and subtangents used to calculate area, *Amer. Math. Monthly*, 109 (December 2002), 900–909.
11. ——, Sums of squares of distances in m-space, *Amer. Math. Monthly*, 110 (June/July 2003), 516–526.
12. ——, Area and arclength of trochogonal arches, *Math Horizons*, (November 2003), 24–30.
13. ——, Isoperimetric and isoparametric problems, *Amer. Math. Monthly*, 111 (February 2004), 118–136.
14. ——, Centroids obtained graphically, *Math. Magazine*, 77 (June 2004), pp. 201–210.
15. ——, A fresh look at the method of Archimedes, *Amer. Math. Monthly*, 111 (June/July 2004), 496–508.
16. ——, Figures circumscribing circles, *Amer. Math. Monthly*, 111 (December 2004), 853–863.
17. ——, Proof without words: Surprising property of hyperbolas, *Math. Magazine*, 79 (2006), 339.

18. ——, Solids circumscribing spheres, *Amer. Math. Monthly*, 113 (June/July 2006), 521–540.

19. ——, Unwrapping curves from cylinders and cones, *Amer. Math. Monthly*, 114 (May 2007), 388–416.

20. ——, The method of punctured containers, *Forum Geometricorum*, 7 (2007), 33–52.

21. ——, New descriptions of conics via twisted cylinders, focal disks, and directors, *Amer. Math. Monthly*, 115 (November 2008), 795–812.

22. ——, The method of sweeping tangents, *Mathematical Gazette*, 92 (No. 525, November 2008), 396–417.

23. ——, A new look at the so-called trammel of Archimedes, *Amer. Math. Monthly*, 116 (February 2009), 115–133.

24. ——, New insight into cycloidal areas, *Amer. Math. Monthly*, 116 (August/September 2009), 598–611.

25. ——, Ellipse to hyperbola: "With this string I thee wed", *Math. Magazine*, 84 (2011), 83–97.

26. ——, Tanvolutes: generalized involutes, *Amer. Math. Monthly*, 117 (October 2010), 701–713.

27. ——, Sums of squares of integers in arithmetic progression, *Mathematical Gazette*, 95 (No. 533, July 2011), 186–196.

28. ——, Complete dissections: converting regions and their boundaries, *Amer. Math. Monthly*, 118 (November 2011), 787–796.

29. ——, New balancing principles applied to circumsolids of revolution, and to n-dimensional spheres, cylindroids, and cylindrical wedges, *Amer. Math. Monthly*, (accepted February, 2012).

30. A. Broman, Holditch's theorem, *Math. Magazine*, 54 (1981), 99–108.

31. B. H. Brown, Conformal and equiareal world maps, *Amer. Math. Monthly*, 42 (1935), 212–223.

32. G. D. Chakerian and P. R. Goodey, Inequalities involving convex sets and their chords, *Annals of Discrete Mathematics*, 20 (1984), 93–101.

33. G. D. Chakerian and M. S. Klamkin, Inequalities for sums of distances, *Amer. Math. Monthly*, 80 (1973), 1009–1017.

34. M. J. Cooker, An extension of Holditch's theorem on the area within a closed curve, *Mathematical Gazette*, 82 (1998), 183–188.

35. R. Courant and H. Robbins, *What is Mathematics?* Oxford University Press, New York, 1996.

36. G. P. Dandelin, Mémoire sur l'hyperboloide de révolution et sur les hexagones de Pascal et de M. Brianchon, *Nouveaux mémoires de l'Académie royale des Sciences et belles-lettres de Bruxelles*, 3 (1826), 1–14. English translation available at http://www.math.ubc.ca∼cass/dandelin.pdf

37. L. E. Dickson, *A History of the Theory of Numbers, Vol. II* (Carnegie Institute, Washington, DC, 1920); reprinted Chelsea Pub., New York, NY, 1952; Dover Pub. Inc., New York, NY, 2005.

38. G. J. Dostor, Questiones sur les nombres, *Archiv der Mathematik und Physik*, 64 (1879), 350–352.

39. W. Feller, *An Introduction to Probability Theory and its Applications, vol. 1*, 2nd ed., John Wiley and Sons, New York, 1957.

40. D. F. Ferguson, Theorems on conics, *Mathematical Gazette*, 31 (1947), 47–49.

41. G. N. Frederickson, *Dissections: Plane and Fancy*, Cambridge University Press, Cambridge, 1997.

42. D. L. Goodstein and J. R. Goodstein, *Feynman's Lost Lecture: The Motion of Planets Around the Sun*. New York: Norton, 1996.

43. J. Gray, Sale of the century?, *Math. Intelligencer*, 21 (3), (1999), 12–15.

44. W. R. Hamilton, The hodograph, or a new method of expressing in symbolic language the Newtonian law of attraction, *Proc. Roy. Ir. Acad.* 3, (1847), 344–353.

45. T. L. Heath, *A History of Greek Mathematics, Vol. I.* Dover, New York, 1981.

46. ——, *A History of Greek Mathematics, Vol. II.* Dover, New York, 1981.

47. ——, *The Works of Archimedes*. Dover, New York, 1953.

48. D. Hilbert and S. Cohn-Vossen, *Geometry and the Imagination*. Chelsea Publishing Co., New York, 1952.

49. H. Holditch, Geometrical theorem, *Quarterly Journal of Pure and Appl. Mathematics*, 2 (1858), 38.

50. M. Hutchings, F. Morgan, M. Ritoré, and A. Ros, Proof of the double bubble conjecture, *Electron. Res. Announc. Amer. Math. Soc.* 6 (2000), 45–49.

51. Roger A. Johnson, *Modern Geometry: an Elementary Treatise on the Geometry of the Triangle and the Circle*, Houghton, Mifflin Co., Boston, 1929.

52. D. Kalman, Archimedes in the 5th dimension, *Math Horizons*, (November 2007), 8–10.

53. J. Kepler, *New Astronomy*; translated by William H. Donahue. Green Lion Press, 2000.

54. G. Loria, *Curve Plane Speciali, Algebrich e Trascendenti; Teoria e Storia,* (Italian). Milano, Ulrico Hoepli, 1930.

55. M. Mamikon, *Kvant* 5 (1977), 10-13 and (1978), 11–17 (Russian).

56. J. H. McKay, The 29th William Lowell Putnam competition, *Amer. Math. Monthly*, 76 (1969), 909–915.

57. L. V. Mirzoyan and Mamikon A. Mnatsakanian, Unusual distribution of flare stars in Pleiades, *International Bulletin of Variable Stars (IBVS),* No. 528 (1971), 1–3.

58. D. S. Mitrinovič, J. E. Pecarič, and V. Volenec, *Recent Advances in Geometric Inequalities.* Kluwer, Dordrecht, 1989.

59. M. A. Mnatsakanian, On the area of a region on a developable surface, *Dokladi Armenian Acad. Sci.* 73 (2) (1981), 97–101. (Russian); communicated by Academician V. A. Ambartsumian.

60. ——, Annular rings of equal area, *Math Horizons*, (November 1997), 5–8.

61. I. Niven, *Maxima and Minima Without Calculus.* Dolciani Mathematical Expositions, no. 6. Math. Assoc. of America, Washington, D.C., 1981

62. G. Salmon, *A Treatise on Conic Sections*, 6th ed. Chelsea, New York, 1960.

63. D. Seiple, E. Boman, and R. Brazier, Mom! There's an astroid in my closet!, *Math. Magazine*, 80 (2007), 104–111.

64. M-K. Siu, On the sphere and cylinder, *College Math. J.*, 15 (1984), 326–328.

65. J. Steiner, *Gesammelte Werke. Band 2*, G. Reimer, Berlin, 1882.

66. H. Steinhaus, *Mathematical Snapshots*, 3rd American Ed., Oxford Univ. Press, New York, 1969.

67. D. J. Struik, *Lectures on Classical Differential Geometry,* 2nd ed. Dover Publications, New York, 1988.

68. A. Todd, Bisecting a triangle, ΠΜΕ *Journal*, 11 (1999), 31–37.

69. L. Withers, Mamikon meets Kepler. email dated April 21, 2009.

70. R. C. Yates, *A Handbook on Curves and Their Properties.* J. W. Edwards, Ann Arbor, 1947.

Index

ABOUT THE AUTHORS

Tom M. Apostol joined the Caltech faculty in 1950 and is Professor of Mathematics, Emeritus. He is internationally known for his books on Calculus, Analysis, and Analytic Number Theory, (translated into 7 languages), and for creating *Project MATHEMATICS!*, a video series that brings mathematics to life with computer animation, live action, music, and special effects. The videos have won first-place honors at a dozen international festivals, and were translated into Hebrew, Portuguese, French, and Spanish. Apostol has published 102 research papers, has written two chapters for the Digital Library of Mathematical Functions (2010), and is coauthor of three texts for the physics telecourse: *The Mechanical Universe ... and Beyond.*

He has received several awards for research and teaching. In 1978 he was a visiting professor at the University of Patras, Greece, and in 2001 was elected a Corresponding Member of the Academy of Athens, where he delivered his inaugural lecture in Greek. In 2012 he was selected to be a Fellow of the American Mathematical Society.

Mamikon A. Mnatsakanian was Professor of Astrophysics at Yerevan State University and Director of the Mathematical Modeling Center of Physical Processes, Armenian Academy of Sciences. As an undergraduate he invented 'Visual Calculus', more fully described in this book. As an astrophysicist he developed a generalized theory of relativity with variable gravitational constant that resolves observational controversies in Cosmology. He also developed new methods that he applied to radiation transfer theory and to stellar statistics and dynamics.

After the 1988 devastating earthquake in Armenia he began seismic safety investigations that brought him to California. After the Soviet Union collapsed, he stayed in the USA where he created assessment problems for the California State Department of Education and UC Davis, and participated in various educational programs, in the process of which he created hundreds of mathematics educational games and puzzles. This work eventually led him to Caltech and *Project MATHEMATICS!*, where he began fruitful collaboration with Tom Apostol. He is the author of 100 scientific papers, 30 of them on mathematics coauthored with Apostol. Website: www.mamikon.com