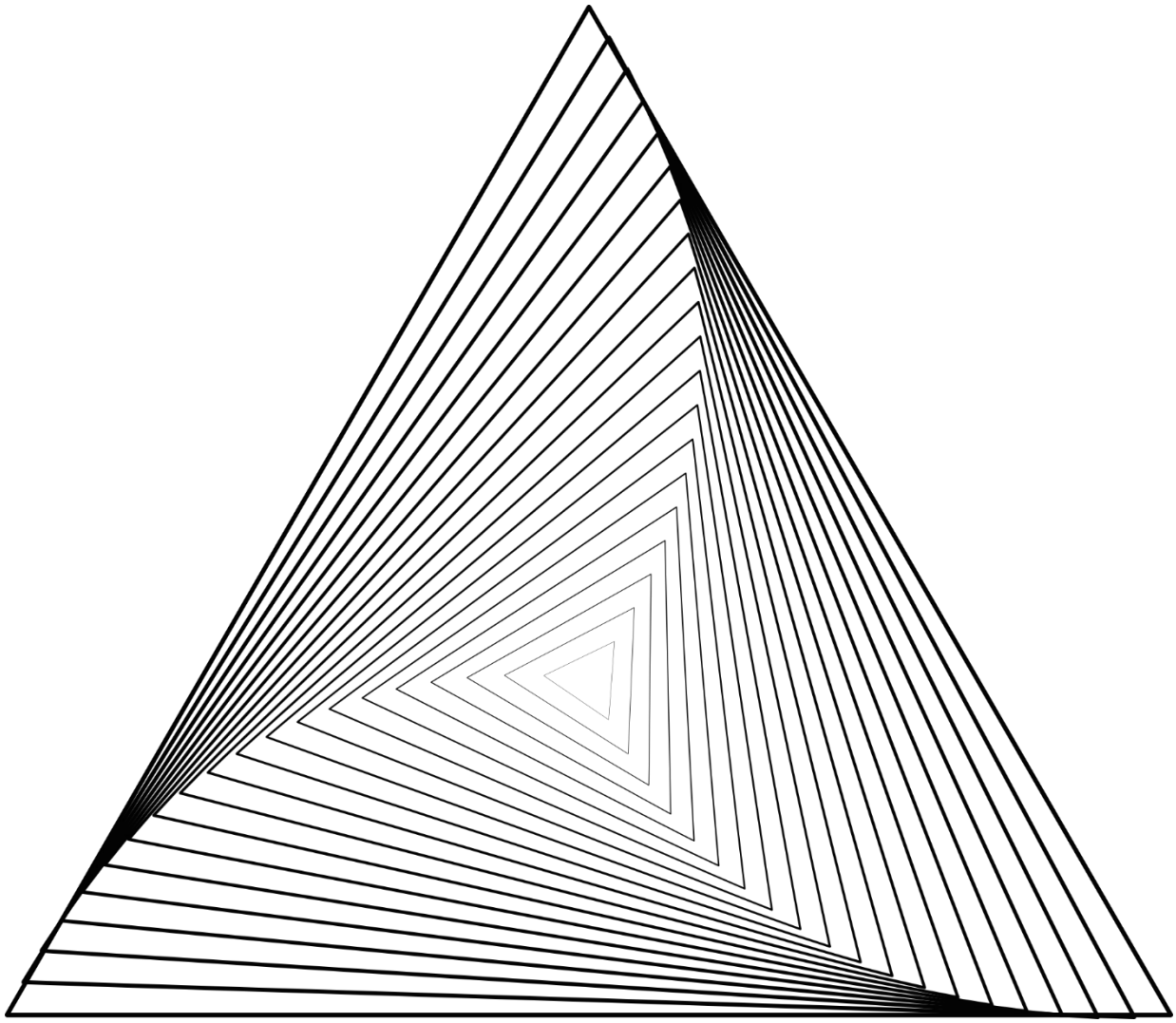
Mathematical Vignettes III



by **Stephen Fratini**

Table of Contents

List of Figures

List of Tables

Preface

God is a mathematician of a very high order and He used advanced mathematics in constructing the universe. – Paul Dirac

"Go down deep enough into anything and you will find mathematics." – Dean Schlicter

This is the third (and maybe not final) book in a series entitled Mathematical Vignettes. Electronic versions of the first two volumes are available free at https://github.com/sfratini33/art-of-managing-things-external/tree/master/free_books.

This book offers in-depth summaries of three topics, i.e., non-Euclidean geometry, topology (including surfaces and metric spaces) and complex analysis. My aim is to introduce readers to each subject, providing ample references for further exploration.

The section on non-Euclidean geometry provides a brief overview of Euclid's parallel postulate and how that led to the eventual development of non-Euclidean geometry. We then go on to cover hyperbolic and elliptic geometries. The prerequisite for this section is a good understanding of Euclidean geometry.

The topology section begins with an examination of surfaces, followed by a discussion on metric spaces, which smoothly transitions into topology as a generalization of these concepts. Topology, a more generalized and abstract form of geometry, is discussed at a level appropriate for upper-level college mathematics students.

The last section covers complex analysis. We start with complex numbers and then continue with complex functions. From there, we discuss continuity, differentiation and integration of complex functions. The concepts of complex series, singularities and residues are also presented. The section is written with the assumption that the reader is familiar with calculus.

Legal

Stephen Fratini
Sole Proprietor of The Art of Managing Things
Eatontown, New Jersey (USA)
Email: sfratini@artofmanagingthings.com or sfratini@outlook.com
LinkedIn: www.linkedin.com/in/stephenfratini

Other books by the author:

- *The Art of Managing Things (2nd edition)*, self-published on Amazon, https://www.amazon.com/Art-Managing-Things-Stephen-Fratini-ebook/dp/B07N4H4YWH/, January 2019.
- *Mathematical Thinking: Exercises for the Mind (2nd Edition)*, self-published on Amazon, https://www.amazon.com/dp/B0CL34FRP1, October 2023.
- *Financial Mathematics (2nd Edition)*, self-published on Barnes and Noble, https://www.barnesandnoble.com/w/financial-mathematics-stephen-fratini/1145166826, March 2023.
- *Math in Art, and Art in Math*, self-published on Amazon, https://www.amazon.com/dp/B091D1F8MB, March 2021.
- *Algebra through Discovery and Experimentation*, self-published on Amazon, https://www.amazon.com/dp/B09B5L9WL5, July 2021.
- *The Struggle Against Chaos*, self-published on Amazon, https://www.amazon.com/dp/B09BLPQ86Q, July 2021.
- *Mathematical Vignettes: Number theory, stochastic processes, game theory, cryptography, linear programming and more*, self-published on Amazon, https://www.amazon.com/Mathematical-Vignettes-stochastic-cryptography-programming-ebook/dp/B0BBP1PBJQ/, August 2022.
- *Learning Math through Puzzles: Number properties, counting, sequences and series, algebra, functions, and mathematical reasoning*, self-published on Amazon, https://www.amazon.com/dp/B0BZFRZP5B, March 2023.
- *Mathematical Vignettes: Volume II: Topics from combinatorial design, magic squares, finite geometry, abstract algebra, error correcting codes, geometric packing problems and much more,* self-published on Amazon, *https://www.amazon.com/dp/B0CM1CLSK8, October 2023.*
- *Shape Up and Solve It!: Learn Geometry Through Puzzles*, self-published on Amazon, https://www.amazon.com/dp/B0CRS7DRWF, January 2024.

Electronic versions of my books are available (free of charge) at

https://github.com/sfratini33/art-of-managing-things-external/tree/master/free_books

1 Introduction

"In mathematics, the art of proposing a question must be held of higher value than solving it."
George Cantor

"The laws of nature are but the mathematical thoughts of God."
Euclid

"A mathematical theory is not to be considered complete until you have made it so clear that you can explain it to the first man whom you meet on the street." – David Hilbert

1.1 Purpose

The purpose of this book is to introduce the reader to three related aspects of mathematics, i.e., non-Euclidean geometry, topology and complex analysis. Further, the intent is to interest the reader in further study. To that end, many references are provided.

1.2 Intended Audience

The intended audience includes mathematically sophisticated readers who are familiar with Euclidean geometry, and calculus, and who wish to expand their knowledge of mathematics.

1.3 Prerequisites

The section on non-Euclidean geometry assumes a knowledge of Euclidean geometry.

The section on topology is mostly self-contained but does assume a level of mathematical sophistication on the part of the reader, e.g., familiarity with proofs and abstract concepts.

The section on complex analysis requires prior knowledge of calculus. If you need to brush-up on your calculus, there are many (free) online resources available, e.g., see the online adaption of APEX Calculus at https://sites.und.edu/timothy.prescott/apex/web/apex.Ptx1.html.

1.4 Outline

The following is a summary of the contexts of this book. The assumption is that the section be read in order but with proper background, it should be possible to skip to any of the three main sections.

- Section 1 is this introduction.
- Section 2 covers non-Euclidean geometry. Section 2.1 discusses Euclid's fifth postulate about parallel lines and how the modification (perhaps "replacement" is a better word) of this postulate has led to various non-Euclidean geometries. Section 2.2 addresses hyperbolic geometry, where given a line and a point not on the line, there exists more than one line through the point and parallel to the given line. In Section 2.3, we discuss elliptic geometry, where there are no parallel lines.
- Section 3 starts with metric spaces and then generalizes the concept to topological spaces. Various concepts such as surfaces, open and closed sets, subspaces, continuous functions, compactness and connectedness are defined and elaborated.

- Section 4 introduces the reader to complex numbers, the complex plane and complex functions (including graphing, continuity, derivatives, integration). Also covered are elementary complex functions (e.g., exponential, log and trigonometric functions), complex series, the classification of singularities, and residue theory.
- At the end of the document, there is a list of acronyms and symbols, a list of references, and an index of terms.

2 Non-Euclidean Geometry

One must be able to say at all times-instead of points, lines, and planes – tables, chairs, and beer mugs. – David Hilbert

"Geometry is the foundation of all painting."
Albrecht Durer

"Geometry existed before the creation. It is co-eternal with the mind of God...Geometry provided God with a model for the Creation." – Johannes Kepler

"Never argue with a $90°$ angle – it's always right." - Anonymous

2.1 Overview

The topic "non-Euclidean geometry" consists of several different geometries based on the axioms of Euclidean geometry with some critical modifications. (See Section 2 of "Shape Up and Solve It!: Learn Geometry Through Puzzles" [1] for a brief introduction to the axioms and basic theorems of Euclidean geometry.) The focus here is on hyperbolic geometry and elliptic geometry, which are defined by modifying the parallel postulate of Euclidean geometry. Affine geometry (not discussed in this book) is another type of non-Euclidean geometry. Affine geometry (or non-metric geometry) is what remains of Euclidean geometry when ignoring (mathematicians often say "forgetting") the metric notions of distance and angle. [2]

The fundamental distinction among the various metric geometries (i.e., geometries that have the concept of distance and angle) lies in the concept of parallel lines. Euclid's fifth postulate, also known as the parallel postulate, is synonymous with **Playfair's postulate** [3]. This postulate (which applies to Euclidean geometry) asserts that in a plane, for any given line ℓ and a point A not on ℓ, there exists exactly one line passing through A that does not intersect ℓ. In hyperbolic geometry, on the other hand, there exist an infinite number of lines through A that do not intersect ℓ. In elliptic geometry, every line through A intersects ℓ.

An alternate way to describe the differences between these geometries (i.e., Euclidean, hyperbolic and elliptic geometries) is to consider two straight lines, indefinitely extended in a two-dimensional plane, which are both perpendicular to a third line (in the same plane):

- In Euclidean geometry, the lines remain at a constant distance from each other (meaning that a line drawn perpendicular to one line at any point will intersect the other line, and the length of the line segment joining the points of intersection remains constant). Such lines are known as parallels.
- In hyperbolic geometry, they "curve away" from each other, increasing in distance as one moves further from the points of intersection with the common perpendicular; these lines are sometimes called ultra-parallels.
- In elliptic geometry, the lines "curve toward" each other and intersect.

The above concepts are illustrated in Figure 1 and Figure 2. The source of the figures is the Wikipedia article “non-Euclidean geometry” [4].

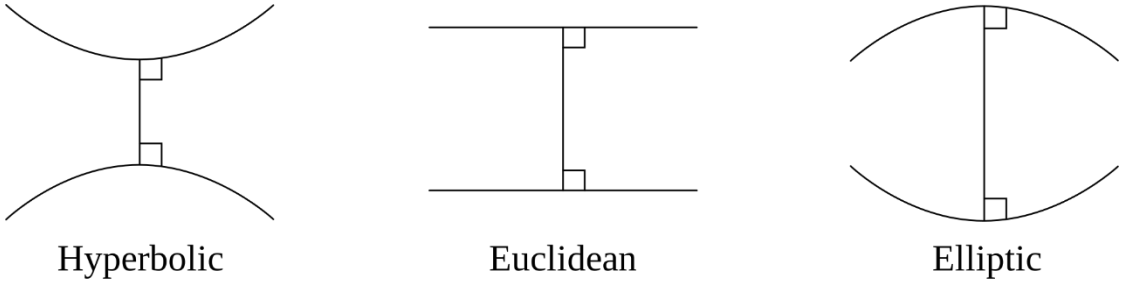


Figure 1. Behavior of lines with a common perpendicular in various geometries

Figure 2 provides some example models for elliptic, Euclidean and hyperbolic geometry.

- Euclidean geometry is modelled by the familiar notion of a "flat plane."
- A simple model for elliptic geometry is a sphere, where lines are "great circles", e.g., the equator or the meridians on a globe.
- The pseudosphere (i.e., a surface with constant negative Gaussian curvature [5]) has the appropriate curvature to model hyperbolic geometry.

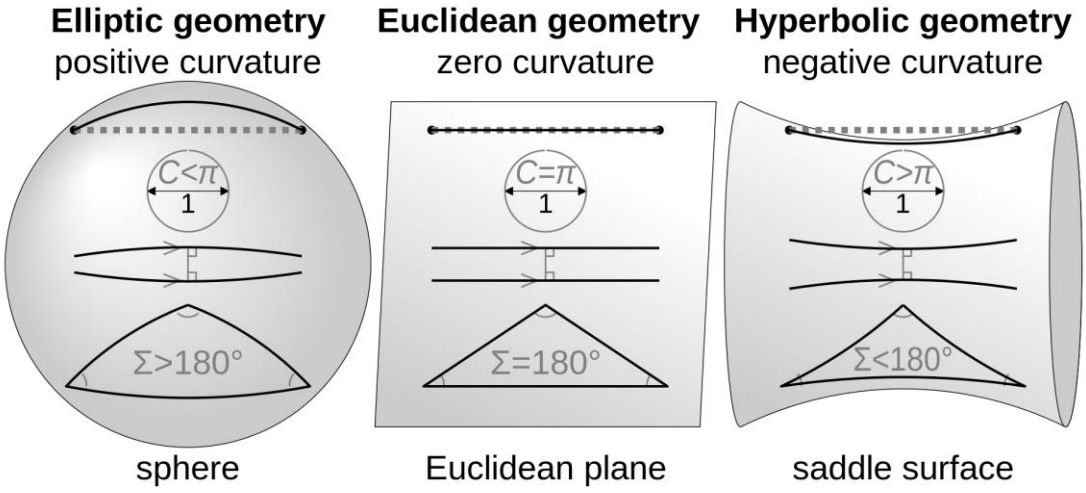


Figure 2. Example models for Elliptic, Euclidean and Hyperbolic geometries

. . .

The following notational conventions are used in what follows:

- Points are indicated by capital letters.
- Angles are indicated by lowercase letters in italics (usually Greek letters with some exceptions).
- Lines are usually represented by script letters, e.g., ℓ, m, n.
- A small square at a vertex of a polygon indicates a right angle ($90°$ or $\frac{\pi}{2}$ radians).

- Figures (in particular, angles) are not always drawn to scale.
- Given points A and B, we use the notation AB to indicate the line AB, the line segment AB and the distance between A and B. The context will make clear which of the three is intended.

2.1.1 Euclid's Fifth Postulate

As noted in the previous section, the modification of Euclid's fifth postulate leads to alternate geometries such as the hyperbolic and elliptic geometries. Before proceeding to discuss these alternate geometries, let's explore Euclid's fifth postulate a bit more.

Translated into English, Euclid's fifth postulate goes as follows:

> If a line segment intersects two straight lines forming two interior angles on the same side that are less than two right angles, then the two lines, if extended indefinitely, meet on that side on which the angles sum to less than two right angles.

Euclid's fifth postulate is illustrated in Figure 3. The two lines ℓ and m form interior angles α and β (via their intersection with line n) such that the sum of the two angles is less than two right angles. The postulate states that ℓ and m will eventually intersect on the side of line n associated with the two said angles. The implication is that if the sum of the two interior angles equals that of two right angles, then the lines ℓ and m do not intersect, i.e., are parallel.

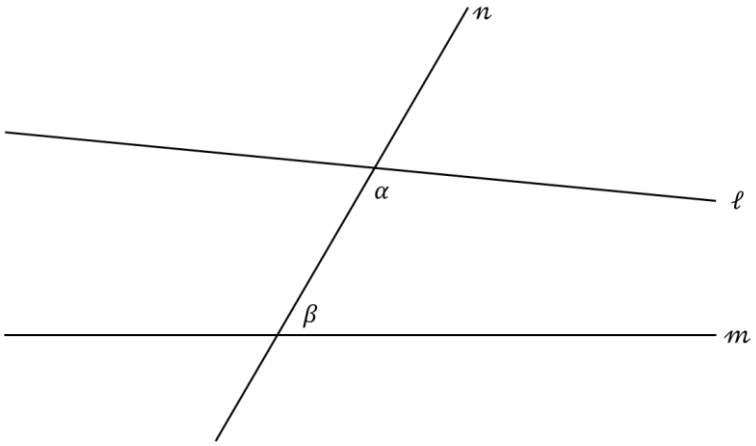


Figure 3. Depiction of Euclid's fifth postulate

There are many equivalent postulates to Euclid's fifth postulate. We already mentioned one, i.e., Playfair's postulate. As Euclid's fifth postulate is the departure point for several non-Euclidean geometries, it is important to know the equivalent statements of the postulate, as listed below [6]:

- Given a line and a point not on the line, there is exactly one line through the point and parallel to the given line. (Playfair's postulate)
- The sum of the angles in every triangle is 180° (triangle postulate).
- There exists a triangle whose angles add up to 180°.
- The sum of the angles is the same for every triangle. [In some non-Euclidean geometries, the sum of different triangles can add to different measures.]

- There exists a pair of similar, but not congruent, triangles.
    - "Congruent" is as defined in Euclid's Elements [12], i.e., "Congruent figures are those that can be made to coincide by superposition. They agree in shape and size, but differ in position."
- Every triangle can be circumscribed by a circle, i.e., the three vertices of a triangle determine a unique circle.
- If three angles of a quadrilateral are right angles, then the fourth angle is also a right angle.
- There exists a quadrilateral in which all angles are right angles. Such a quadrilateral is known as a rectangle.
- There exists a pair of straight lines that are at a constant distance from each other.
- Two lines that are parallel to the same line are also parallel to each other.
- In a right triangle (i.e., a triangle with a right angle), the square of the hypotenuse equals the sum of the squares of the other two sides (Pythagorean theorem).
- The law of cosines [7] holds true (a generalization of the Pythagorean theorem that holds for all triangles, not just right triangles).
- There is no upper limit to the area of a triangle (Wallis' axiom).
- The summit angles of the Saccheri quadrilateral [8] are right angles (the following section provides further details on this topic).
- If a line intersects one of two parallel lines, both of which are coplanar with the original line, then it also intersects the other (Proclus' axiom).

2.1.2 Saccheri Quadrilateral

A **Saccheri quadrilateral** has two equal sides perpendicular to its base. It is named after Giovanni Gerolamo Saccheri, who used it extensively in his 1733 book "Euclid freed of every flaw", an (unsuccessful) attempt to prove the parallel postulate using the method reductio ad absurdum (i.e., proof by contradiction).

For a Saccheri quadrilateral $ABCD$, the legs AD and BC are equal in length and perpendicular to the base AB, see Figure 4. The top side (CD in the figure) is called the summit and the angles at vertices C and D are called the summit angles.

The advantage of using Saccheri quadrilaterals when considering the parallel postulate is that they present three mutually exclusive options, i.e., the summit angles are right angles, obtuse angles, or acute angles (see Figure 4). The right angle case applies to Euclidean geometry. The acute angle case applies to (actually defines) hyperbolic geometry. In elliptic or spherical geometry, the summit angles are always obtuse (think of drawing a quadrilateral on a globe).

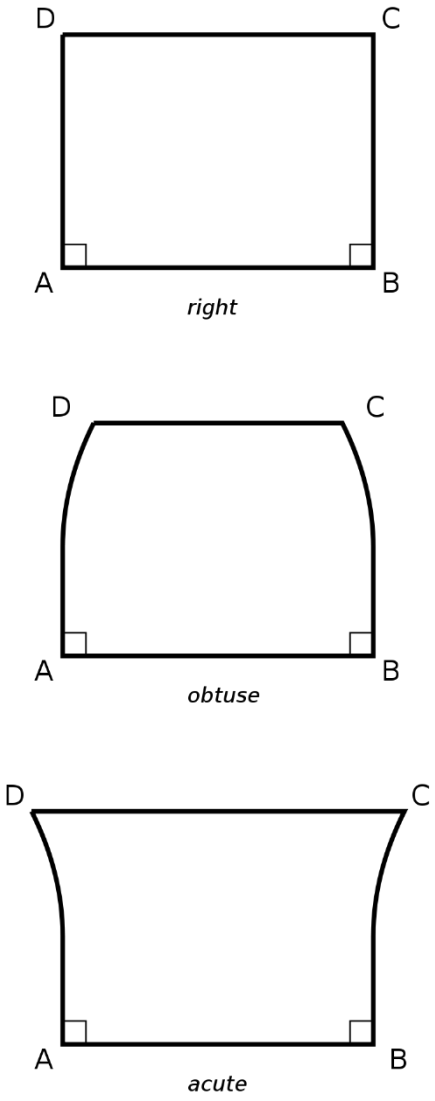


Figure 4. Three cases for summit angles in a Saccheri quadrilateral

2.1.3 History of the Parallel Postulate

This section is a shortened and edited version of the Wikipedia article entitled “Parallel postulate” [6]. The main points are as follows:

- mathematicians struggled for millennia with Euclid’s fifth postulate (parallel postulate) before agreeing that it is in fact a postulate (and not a theorem to be proved from other postulates)
- the parallel postulate can be modified to allow for other types of consistent geometries.

From its inception, the parallel postulate (also known as Euclid’s fifth postulate) came under question as being provable from the other four postulates stated by Euclid, and thus, not a postulate but a theorem.

Euclid’s postulates are as follows:

I. A straight line segment can be drawn joining any two points, i.e., two points uniquely determine a line.

II. Any straight line segment can be extended indefinitely in a straight line.

III. Given any straight line segment, a circle can be drawn having the segment as radius and one endpoint as center.

IV. All right angles are congruent.

V. The parallel postulate which we described in the previous section.

For more than two thousand years, many attempts were made to prove the parallel postulate using Euclid's first four postulates – all to no avail. A proof of the parallel postulate was so highly sought after because unlike the first four postulates, the parallel postulate is not self-evident. If the order in which the postulates were listed in Euclid’s Elements [9] is significant, it indicates that Euclid included this postulate only when he realized he could not prove it or proceed without it. Many attempts were made to prove the fifth postulate from the other four, with many of them being accepted as proofs for long periods until a mistake was found. Invariably the mistake was assuming some “obvious” property which turned out to be equivalent to the fifth postulate (e.g., Playfair's postulate). Today, over two thousand two hundred years later, Euclid's fifth postulate properly remains a postulate and not a theorem.

. . .

Proclus (410–485) wrote a commentary on Euclid’s Elements where he discussed “attempted proofs” to deduce the fifth postulate from the other four; in particular, he notes that Ptolemy had produced a false proof. Proclus then goes on to give a false proof of his own. However, he did state a postulate (now known as Playfair’s postulate) which is equivalent to the fifth postulate.

Ibn al-Haytham (Alhazen) (965–1039), an Arab mathematician, attempted to prove the parallel postulate using a technique known as “proof by contradiction”. While his proof was invalid, he did introduce the concept of motion and transformation into geometry. [**Author’s Remark**: It is not uncommon in mathematics to discover new concepts and theories while trying to solve another problem.]

Nasir al-Din al-Tusi (1201–1274) wrote detailed critiques of the parallel postulate. Nasir al-Din attempted to derive a proof by contradiction of the parallel postulate. Most notably, he also considered the cases of what are now known as elliptical and hyperbolic geometry, though he ruled out both of them.

Nasir al-Din's son, Sadr al-Din (also known as Pseudo-Tusi), authored a book on the subject in 1298, based on his father's later thoughts, which presented one of the earliest arguments for a non-Euclidean hypothesis equivalent to the parallel postulate. Further, he revised the Euclidean system of postulates and theorems to align with his non-Euclidean version of the parallel postulate. His work was published in Rome (1594) and influenced European geometers such as Giordano Vitale (1633–1711), Girolamo Saccheri (1667–1733) and Johann Lambert (1728–1777). Lambert proved the non-Euclidean result that the sum of the angles in a triangle increases as the area of the triangle decreases.

In the nineteenth century, mathematicians more fully considered alternatives to Euclidean geometry. In this period, logically consistent non-Euclidean geometries were developed. Nikolai

Ivanovich Lobachevsky (in 1829) and János Bolyai (in 1831) independently defined what we now consider to be hyperbolic geometry.

It should be added that famous and prolific mathematician Carl Friedrich Gauss (1777-1855) may have been the first to fully accept the possibility of non-Euclidean geometry, but he did not formally publish his work on this topic. In a letter written at Gottingen on 8 November 1814 to F. A. Taurinus, Gauss made the following statement. The letter was translated from German to English in the passage below, see Section 22 of the book by Wolfe [10].

> **The assumption that the sum of the three angles is less than 180° leads to a curious geometry, quite different from ours (the Euclidean), but thoroughly consistent, which I have developed to my entire satisfaction**, so that I can solve every problem in it with the exception of the determination of a constant, which cannot be designated a priori. The greater one takes this constant, the nearer one comes to Euclidean Geometry, and when it is chosen infinitely large the two coincide. The theorems of this geometry appear to be paradoxical and, to the uninitiated, absurd; but calm, steady reflection reveals that they contain nothing at all impossible. For example, the three angles of a triangle become as small as one wishes, if only the sides are taken large enough; yet the area of the triangle can never exceed a definite limit, regardless of how great the sides are taken, nor indeed can it ever reach it. **All my efforts to discover a contradiction, an inconsistency, in this Non-Euclidean Geometry have been without success,** and the one thing in it which is opposed to our conceptions is that, if it were true, there must exist in space a linear magnitude, determined for itself (but unknown to us). But it seems to me that we know, despite the say-nothing word-wisdom of the metaphysicians, too little, or too nearly nothing at all, about the true nature of space, to consider as absolutely impossible that which appears to us unnatural. If this Non-Euclidean Geometry were true, and it were possible to compare that constant with such magnitudes as we encounter in our measurements on the earth and in the heavens, it could then be determined a posteriori. Consequently, in jest, I have sometimes expressed the wish that the Euclidean Geometry were not true, since then we would have a priori an absolute standard of measure.

The resulting geometries were later developed by Lobachevsky, Bernhard Riemann and Henri Poincaré into hyperbolic geometry and elliptic geometry. The independence of the parallel postulate from Euclid's other axioms was finally demonstrated by Eugenio Beltrami in 1868.

2.2 Life without the Parallel Postulate

In this section, we present some results that can be derived from Euclid's first four postulates. These results will be used in our study of hyperbolic geometry in the next section. Propositions 1 through 28 in Book I of Euclid's Elements only require the first four postulates, and so, we assume these results. [**Author's Remark**: We do not state the said 28 propositions (i.e., theorems) here since they are available in a free publication, see "Project Gutenberg's First Six Books of the Elements of Euclid" [12]. There is also an online version of Euclid's Elements hosted by Clark University [13]. In what follows, we will reference such proposition using the naming scheme from the said publication, e.g., "Book III, Prop. IX.". This will allow the reader to do an easy search to find the proposition.]

Further, there are some very basic assumptions and axioms that were missed by Euclid. These assumptions and axioms were identified by famous mathematician David Hilbert, see the Wikipedia article on Hilbert's axioms [14]. We will make use of these items as needed.

We also need a few definitions related to the Saccheri quadrilateral $ABCD$. In Figure 5, AD and CB are the arms of the quadrilateral, the angles at D and C are known as summit angles, DC is the summit and AB is the base. The arms in a Saccheri quadrilateral are of equal length. As we shall prove, the summit angles are not obtuse (i.e., not greater than $90°$) if we remove the parallel postulate.

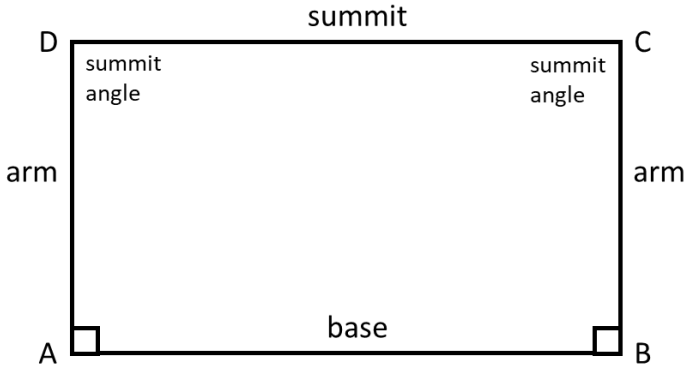


Figure 5. Terminology for Saccheri quadrilateral

The theorems in this section apply to any geometry that can be built on the axioms and assumptions of Euclidean geometry **minus** the parallel postulate. In particular, these theorems apply to hyperbolic geometry but not to elliptic geometry (which depends on a different and more complex set of axioms).

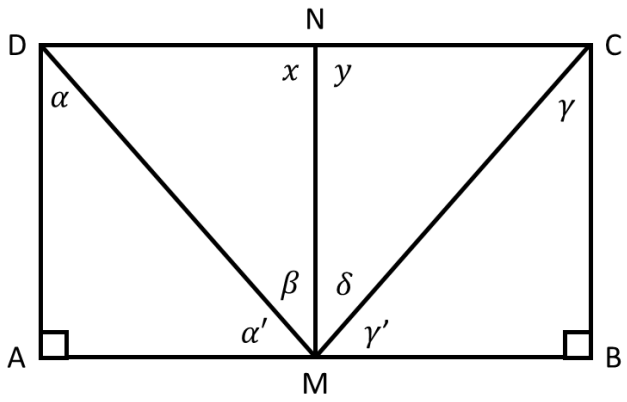
Theorem 1. The summit angles of a Saccheri quadrilateral are equal.

Proof: Referring to Figure 5, triangles ABC and BAD are congruent by the Side-Angle-Side (SAS) triangle congruence principle (Book I, Prop. IV of Euclid's Elements). Thus, $AC = BD$ (this is shorthand for saying that segments AC and BD are of equal length).

By the Side-Side-Side (SSS) triangle congruence principle (Book I, Prop. VIII of Euclid's Elements), triangles ADC and BCD are congruent, which implies $\angle D = \angle C$. ■

Theorem 2. The line segment between the midpoints of the base and summit of a Saccheri quadrilateral is perpendicular to the base and the summit.

Proof: In the figure below, N is the midpoint of line segment DC, and M is the midpoint of line segment AB. By the definition of a Saccheri quadrilateral, we know that $AD = BC$.



By the SAS triangle congruence principle, triangles ADM and BCM are congruent, and thus, $DM = CM, \alpha' = \gamma'$ and $\alpha = \gamma$.

By the SSS triangle congruence principle, triangles DNM and CNM are congruent. So, angles x and y are equal, and since they comprise a straight line, $x = y = 90°$. Further, $\beta = \delta$.

Since $\alpha' + \beta = \gamma' + \delta$ and the four angle comprise a straight line, we have that $\alpha' + \beta = \gamma' + \delta = 90°$.

Thus, MN is perpendicular to the summit DC and the base AB. ■

Theorem 3. The base and summit of a Saccheri quadrilateral are parallel.

Proof: This follows from Book I, Prop. XXVII of Euclid's Elements and the result of Theorem 2. ■

Note: Euclid's Elements uses the term "right line" to mean what is more usually called a "straight line" in present terminology.

In Euclidean geometry, we know that the sum of the angles of a triangle add to $180°$. However, when we remove the parallel postulate, we can only prove the following theorem.

Theorem 4. The sum of the angles of a triangle are less than or equal to $180°$.

Proof: See Theorem 31 in Chapter III of Gans [11].

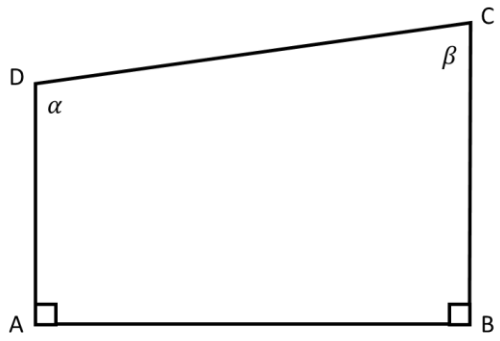
Theorem 5. The summit angles of a Saccheri quadrilateral are not obtuse, i.e., they are either acute or right angles.

Proof: Consider the Saccheri quadrilateral $ABCD$ in Figure 5. If the summit angles were obtuse, the sum of angles of the quadrilateral would be greater than $360°$. The sum of the angles in triangles ABC and ADC is equal to the sum of the angles in quadrilateral $ABCD$, and thus (under our assumption) would be greater than $360°$. Thus, the sum of the angles in either triangle ABC or ADC would need to surpass $180°$, but this contradicts Theorem 31. Therefore, the summit angles cannot be obtuse. ■

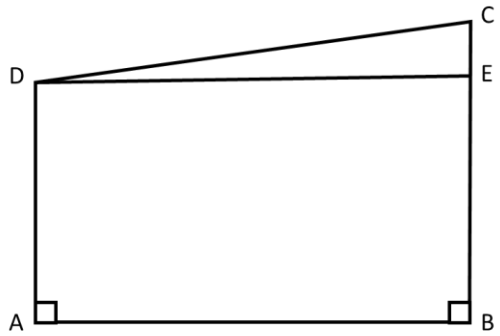
The following theorem applies to quadrilaterals that have right angles at their base but are not necessarily Saccheri quadrilaterals.

Theorem 6. Given a quadrilateral with right angles at its base. If the edges (arms) extending from the base are unequal, then so are the associated summit angles, and conversely. Further, the greater summit angle lies opposite the greater arm.

The quadrilateral $ABCD$ in the figure below meets the conditions of the theorem. The summit angles are α and β, the arms are AD and BC, and the base is AB. The theorem says that if $\alpha > \beta$ then the length of arm AD is less than the length of arm BC.



Proof: Starting with the figure above, assume that $AD < BC$. Select a point E such that $AD = BE$ (as shown in the figure below). Quadrilateral $ABED$ is, by definition, a Saccheri quadrilateral. By Theorem 1, $\angle ADE = \angle BED$.



Since line segment DE subdivides $\angle ADC$, we have that $\angle ADC > \angle ADE$ (this follows from one of the unstated assumptions in Euclid's Elements).

Since $\angle BED$ is an exterior angle to triangle CED (at point E), we have that $\angle BED > \angle BCD$ (by Book I, Prop. XVI of Euclid's Elements).

Putting the above results together, gives us the desired result, i.e., $\angle ADC = \angle BED > \angle BCD$. So, if the summit angles are unequal, the angle of greater measure lies opposite the arm of greater measure.

Going in the other direction, assume that summit angles of quadrilateral $ABCD$ are unequal, i.e., (without loss of generality) assume $\angle ADC > \angle BCD$. The arms AD and BC cannot be equal since this would imply that the summit angles are equal by Theorem 1. Thus, $AD > BC$ or $BC > AD$. The former inequality cannot be true, since it would imply that $\angle BCD > \angle ADC$ (this follows from the first part of this theorem). So, we must conclude that $BC > AD$. Thus, the arms are unequal, with the greater one lying opposite the greater summit angle. ■

2.3 Hyperbolic Geometry

2.3.1 Basic Concepts

Hyperbolic geometry is based on the first four axioms of Euclidean geometry, plus the negation of the parallel postulate. Given the many equivalent statements to the parallel postulate, there are many ways to state its negation. We take the approach of negating the statement concerning Saccheri quadrilateral.

Recall that the following statement is equivalent to the parallel postulate:

> The summit angles of the Saccheri quadrilateral are right angles.

From Theorem 5 (which only assumes the first four of Euclid's axioms), we know that the summit angles of the Saccheri quadrilateral are either acute or right angles. So, if we assume the summit angles are acute, then we effectively negate the parallel postulate. Thus, we have the following postulate (axiom) for hyperbolic geometry. [**Author's Remark**: This looks so simple but it took almost 2000 years after the work of Euclid to come to this conclusion.]

> **Hyperbolic parallel postulate**: The summit angles of a Saccheri quadrilateral are acute.

Figure 6 is an alternative drawing of the Saccheri quadrilateral in the case of hyperbolic geometry, with emphasis on the summit angles being acute. We will, however, continue to use straight lines for most of the subsequent illustrations in this section.

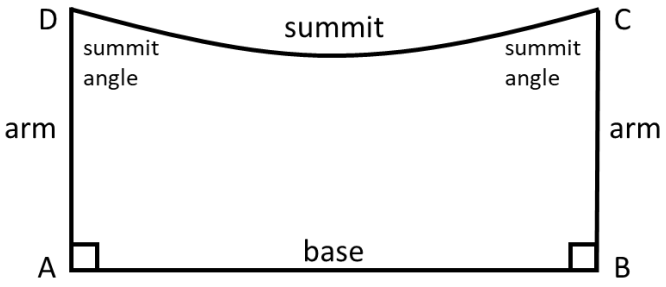


Figure 6. Saccheri quadrilateral in hyperbolic geometry

The negation of the parallel postulate (via the hyperbolic parallel postulate stated above) immediately gives us several facts based on the negation of the statements that are equivalent to the parallel postulate. We list a few of these facts below:

- Given a line and a point not on the line, there exists more than one line through the point and parallel to the given line. [Later on, we will prove that given any line and any point not on it, there exists infinitely many lines through the point which are parallel to the line, and have a common perpendicular with it.]
- There exist triangles whose angle sums are different.
- All similar triangles are congruent.
- There does not exist any pair of straight lines that are at constant distance from each other.
- There does not always exist a circle passing through three given noncollinear points.

. . .

Closely related to the Saccheri quadrilateral is the **Lambert quadrilateral**. In particular, a Lambert quadrilateral (also known as Ibn al-Haytham–Lambert quadrilateral), is a quadrilateral in which three of its angles are right angles. See the example of a Lambert quadrilateral in Figure 7. The summit line segment is drawn as a curve to emphasize that the angle at C is acute (this drawing convention will not be used in subsequent figures).

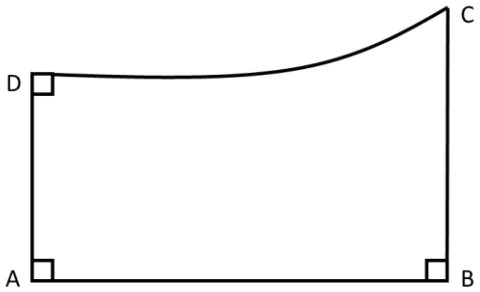
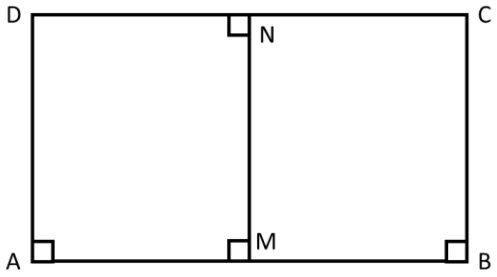


Figure 7. Example of a Lambert quadrilateral

A Lambert quadrilateral can be constructed from a Saccheri quadrilateral by joining the midpoints of the base and summit of the Saccheri quadrilateral (recall Theorem 2). This line segment is perpendicular to both the base and summit and so each half of the Saccheri quadrilateral is a Lambert quadrilateral. The proof of the following theorem makes use of two Lambert quadrilaterals that comprise a single Saccheri quadrilateral.

Theorem 7. In a Saccheri quadrilateral, the summit is of length greater than the base, and the line segment joining the midpoints of the summit and base is shorter than either arm.

Proof: Consider the Saccheri quadrilateral $ABCD$ in the figure below, where point M is the midpoint of the base and point N is the midpoint of the summit. By Theorem 2, the line segment MN is perpendicular to the summit and the base (as indicated in the figure).



Since $ABCD$ is a Saccheri quadrilateral (by assumption), then the angles at D and C are acute. Applying Theorem 6 to Lambert quadrilateral $AMND$ (viewing AD and MN as arms), we have that $MN < AD$. Applying Theorem 6 to Lambert quadrilateral $MBCN$ (viewing BC and MN as arms), we have that $MN < BC$. This proves the second part of the proof.

Applying Theorem 6 to Lambert quadrilateral $AMND$ again but viewing AM and DN as arms, we have that $DN > AM$. Applying Theorem 6 to Lambert quadrilateral $MBCN$ again but viewing BM and CN as arms, we have that $CN > BM$. Adding these two inequalities, we get

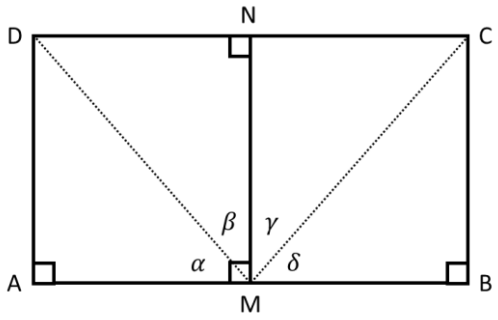
$$DN + CN > AM + BM$$

which implies $DC > AB$, i.e., the length of the summit is greater than the length of the base of Saccheri quadrilateral $ABCD$. ■

In the above proof, lines AB and CD are parallel by Book I, Prop. XXVIII of Euclid's Elements since they have a common perpendicular. Further, AD and MN are unequal distances between lines AB and CD. Hence, we have shown that parallel lines exist which are not equidistant from one another which confirms one of the negations of the parallel postulate.

Theorem 8. The non-right angle in a Lambert quadrilateral is acute and each side adjacent to the non-right angle is longer than the opposite side.

Proof: In the figure below, $MBCN$ is a Lambert quadrilateral with its non-right angle at point C. Extend line segment NC to a point D such that $DN = NC$. Let A be the perpendicular projection of D onto line MB.



Triangles MND and MNC are then congruent by the SAS triangle congruence principle which implies that $MD = MC$, and $\beta = \gamma$ which, in turn, implies $\alpha = \delta$.

Triangles MDA and MCB are congruent by Book I, Prop. XXVI of Euclid's Elements. Thus, $AD = BC$, and $ABCD$ is (by definition) a Saccheri quadrilateral. The summit angle $\angle C$ of $ABCD$ is acute by the hyperbolic parallel postulate, and thus, the non-right angle of the Lambert quadrilateral $MBCN$ is acute.

By Theorem 6, $BC > MN$ and $NC > MB$. ■

2.3.2 Parallel Lines with a Common Perpendicular

If two lines have a common perpendicular, then they are parallel by Book I, Prop. XXVII of Euclid's Elements. This is true for Euclidean and hyperbolic geometry. If two lines had two (or more) common perpendiculars, the two lines would bound a Lambert quadrilateral with all right angles, but this contradicts Theorem 8. So, we have the following result. By similar reasoning, there are no rectangles in hyperbolic geometry.

Theorem 9. Two parallel lines can have at most one common perpendicular.

As we shall see later, it is possible (in hyperbolic geometry) for two parallel lines to not have a common perpendicular.

Book I, Prop. XXVII and Prop. XXVIII of Euclid's Elements states that two lines are parallel if a transversal (i.e., a distinct line that intersects two parallel lines) creates equal alternate interior angles, or equal corresponding angles, respectively.

For example, consider the lines m and n, and transversal ℓ in Figure 8. Prop. XXVII tells us that lines m and n are parallel if alternate interior angles x and y are equal. Prop. XXVIII tells us that lines m and n are parallel if corresponding angles y and z (or v and w) are equal.

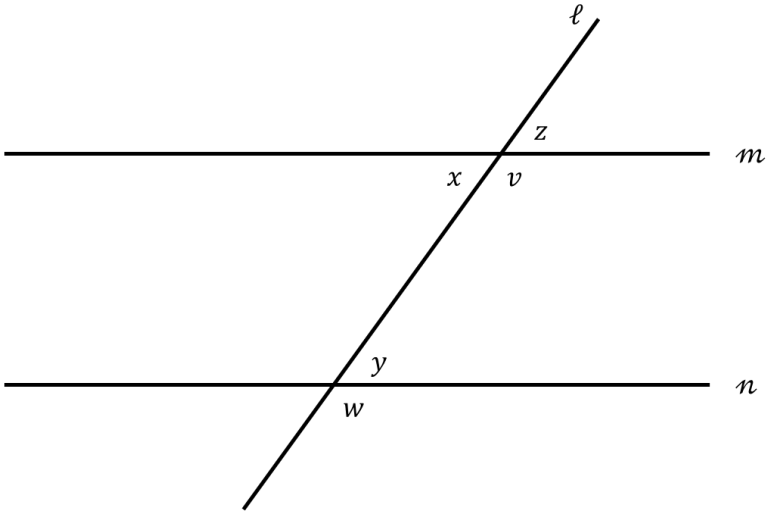


Figure 8. Alternate interior angles and corresponding angles

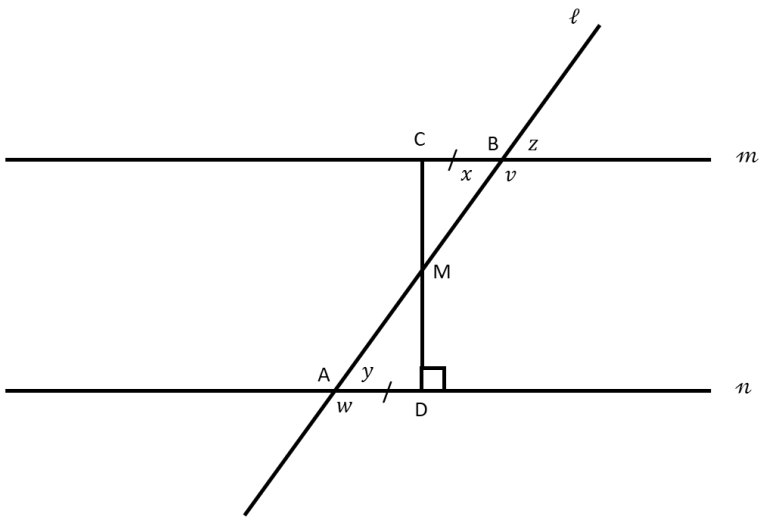
In hyperbolic geometry, a transversal (of two lines m and n) that gives rise to equal alternate interior angles, or equal corresponding angles implies m and n (in addition to being parallel) also have a common perpendicular.

Theorem 10. If two lines have a transversal which intersects the lines to form equal alternate interior angles or equal corresponding angles, then the two lines are parallel and have a common perpendicular.

Proof: In either case, lines are parallel by Book I, Prop. XXVII and XXVIII of Euclid's Elements.

If alternate interior angles (or corresponding angles) are right angles, then we are done, i.e., the two lines have a common perpendicular and by Book I, Prop. XXVII of Euclid's Elements, the two lines are parallel.

In the case where the alternate interior angles are acute, i.e., $x = y < 90°$ in the figure below (or equivalently, the corresponding angles are obtuse). Let M be the midpoint between on line segment AB, where A and B are the points of intersection of the transversal ℓ with the lines m and n. Draw a perpendicular line segment from M to line n (intersecting at point D). Select point C on line m such that $CB = AD$. (At this point, we have not established that C, M and D are collinear.) By the SAS triangle congruence principle, triangles ADM and BCM are congruent which implies that $\angle BCM = \angle ADM = 90°$. Thus, C, M and D are collinear, and CD is a common perpendicular to lines m and n.



If x and y are obtuse angles, then the other two alternate interior angles are acute, and the preceding argument applies. ■

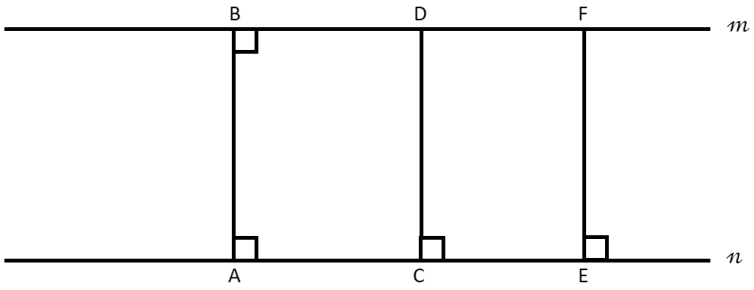
The converse of Theorem 10, which we state below without proof, is also true.

Theorem 11. If two lines have a common perpendicular, then there exists a transversal which intersects the lines such that equal alternate interior angles (and equal corresponding angles) are formed. Further, the only transversals with this property are those which go through the midpoint of the common perpendicular to the two lines.

In the case of two parallel lines with a common perpendicular, the following theorem gives us additional information concerning the distance between the two lines. In the case of Euclidean geometry, this is not true, i.e., the distance between parallel lines is constant.

Theorem 12. The distance between two parallels with a common perpendicular is smallest when measured along the common perpendicular. The distance between the parallel lines increases as one moves further away from the common perpendicular.

Proof: In the figure below, line segment AB is the common perpendicular to parallel lines m and n. Choose a point D (different from B) on line m and draw a perpendicular from D down to line n (with endpoint labeled as C). By definition $ABDC$ is a Lambert quadrilateral. By Theorem 8, $\angle BDC$ is acute and $DC > AB$. Hence, the distance from m to n is less when measured along the common perpendicular than along any other perpendicular from m to n. Using a similar argument, we can show that the distance from n to m is less when measured along the common perpendicular than along any other perpendicular from n to m.

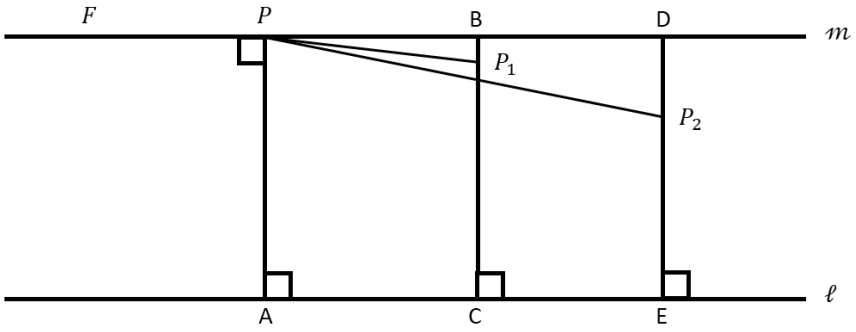


To prove the second part of the theorem, choose any point F on m such that D is between B and F. Draw a perpendicular from F down to line n (with endpoint labeled as E). By definition, $ABFE$ is a Lambert quadrilateral. By Theorem 8, $\angle DFE$ is acute. Further, $\angle CDF$ is obtuse, since we previously established that $\angle BDC$ is acute. Thus, $\angle CDF > \angle BDC$ and by Theorem 6, $FE > CD$, i.e., the distance between the parallel lines increases as we get further away from the common perpendicular. ■

In hyperbolic geometry, unlike Euclidean geometry, there are an infinite number of lines through point P parallel to line ℓ, where P is not on ℓ. Regarding the proof of the following theorem, this is a good time to repeat and emphasize that figures are not drawn to scale, especially with regard to angles.

Theorem 13. Given any line ℓ and any point P not on ℓ, there exists infinitely many lines through P which are parallel ℓ and that have a common perpendicular ℓ.

Proof: Draw a perpendicular from P down to line ℓ (with endpoint A), as shown in the figure below. Position line m such that $\angle FPA = 90°$. By Theorem 10, m is one line through P which is parallel to ℓ and which has a common perpendicular with ℓ (i.e., line AP).



We construct another parallel to ℓ which is distinct from m. Let B be a point on m to the right of P. Draw a perpendicular from B down to line ℓ (with endpoint C). By Theorem 12, we have that $CB > AP$. Select point P_1 on line segment BC such that $AP = CP_1$. By definition, APP_1C is a Saccheri quadrilateral and by Theorem 2 and Theorem 3, lines ℓ and PP_1 are parallel lines with a common perpendicular, i.e., the line defined by the midpoints of AC and PP_1. Thus, we have constructed a second line that is parallel to ℓ and which has a common perpendicular with ℓ.

Next, select a point D on m and to the right of B. Using an argument similar to that in the above paragraph, we can construct another line PP_2 which is parallel to ℓ and has a common perpendicular with ℓ, i.e., the line defined by the midpoints of AE and PP_2. Since $AP = CP_1 = EP_2$,

and the distances at different points between two lines cannot be equal (by Theorem 12), it must be that PP_1 and PP_2 are distinct lines.

So, for each point on m to the right of P there corresponds a unique line which goes through P, is parallel to ℓ, and has a common perpendicular with ℓ. Thus, there are an infinite number of lines through P that are parallel to ℓ and which have a common perpendicular with ℓ. ■

Using the construction in the above proof, the farther to the right of point P that select a point P_n, the smaller the angle $\angle APP_n$. Since there is no furthest position to the right of P, none of these parallels PP_n is closest to line AP in the sense of making a smaller angle with it than do all the others. Clearly, the same is true if we select points on line m to the left of P. It turns out that this property is true regardless of the construction used to determine the various parallels to line ℓ going through point P. We capture this result in the following theorem.

Theorem 14. Let ℓ be any line, P any point not on it, and A the perpendicular projection of P onto ℓ. Among all the lines through P, parallel to ℓ and which have a common perpendicular with ℓ, there is none that lies closest to line AP in the sense of making a smaller angle with it than do all the others.

Proof: See the proof of Theorem 41 in Chapter III Section 7 of the book by Gans [11].

2.3.3 Triangles

One of the equivalents to the parallel postulate is the following statement:

> The sum of the angles in every triangle is $180°$.

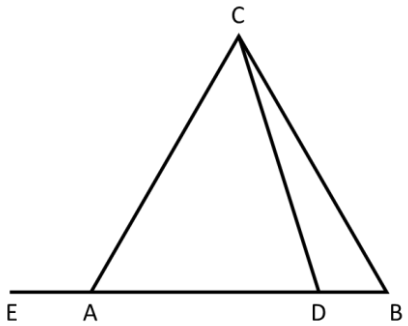
The hyperbolic parallel postulate negates an equivalent of the above statement. Thus, for hyperbolic geometry the sum of the angles in a triangle is not $180°$. By Theorem 4, the sum of the angles of a triangle are not greater than $180°$. Thus, we have proven the following theorem for hyperbolic geometry.

Theorem 15. The sum of the angles of a triangle is less than $180°$.

We also have the following theorem concerning the sum of the angles in a triangle.

Theorem 16. There are triangles with angle-sums arbitrarily close to $180°$.

Proof: Take any triangle ABC (e.g., the one in the figure below) and a variable point D between vertices A and B.

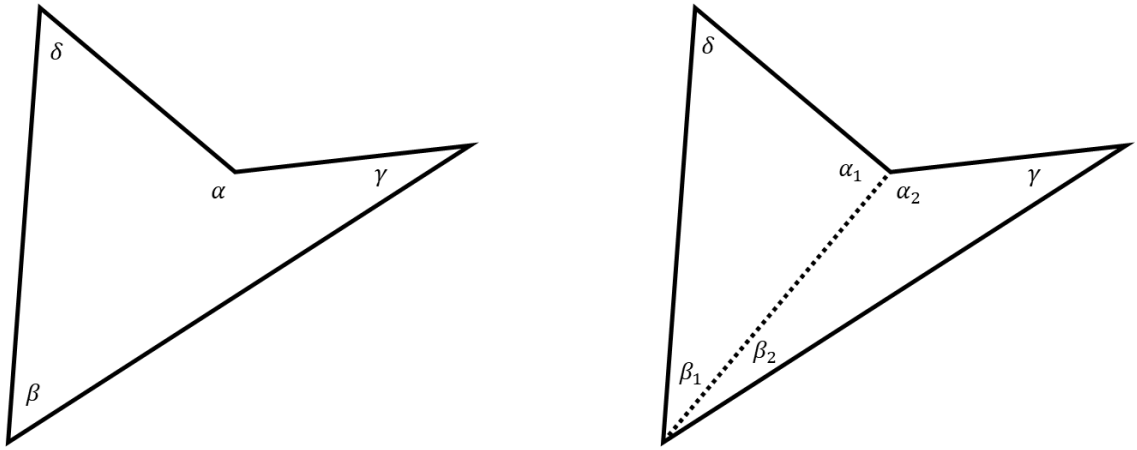


As D approaches A, the measure of $\angle ADC$ approaches (from below) the measure of $\angle EAC$, and the measure of $\angle ACD$ approaches 0. Thus, the sum of the angles of triangle ADC can be made arbitrarily close to the measure of $\angle EAC + \angle CAB$, which is $180°$. ■

The following theorem implies that there are no squares or rectangles in hyperbolic geometry.

Theorem 17. The sum of the angles of a simple quadrilateral (i.e., non-self-intersecting) is less than $360°$.

Proof: Consider the quadrilateral on the left of the figure below. Draw a line to dissect the quadrilateral into two triangles (shown on the right of the figure). By Theorem 15, $\delta + \alpha_1 + \beta_1 < 180°$ and $\gamma + \alpha_2 + \beta_2 < 180°$. Adding the two inequalities, we get $\alpha + \beta + \gamma + \delta < 360°$.

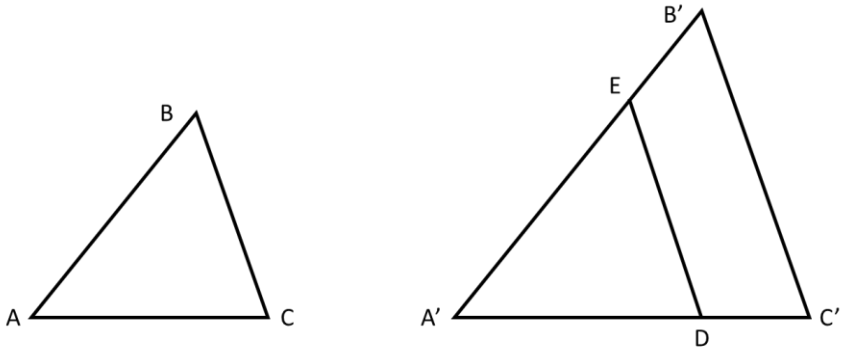


A similar dissection into two triangles can be performed for any simple quadrilateral. ■

In Euclidean geometry, similar triangles can be of different sizes. For example, we can have an equilateral triangle that is the size of a coin and another (similar) equilateral triangle that is the size of the solar system. In hyperbolic geometry, this is not possible.

Theorem 18. If the angles of two triangles are equal, then the triangles are congruent (i.e., their respective sides are equal).

Proof: In the figure below, assume that triangles ABC and $A'B'C'$ have equal angles.



Assume that $AB < A'B'$. Select D on $A'C'$ such that $AC = A'D$, and select E on $A'B'$ (on the same side of A' as B') such that $AB = A'E$. By the SAS triangle congruence principle, triangles ABC and $A'ED$ are congruent. Thus, $\angle ACB = \angle A'DE = \angle C'$ and $\angle ABC = \angle A'ED = \angle B'$. Further, E is between A' and B'; otherwise, an exterior angle of a triangle (in this case $\angle A'ED$) would equal to an opposite interior angle of the triangle (in this case $\angle B'$) which contradicts Book I, Prop. XVI of Euclid's Elements.

Since $A'DE + \angle EDC' = 180°$ and $\angle A'DE = \angle C'$, we have that $\angle EDC' + \angle C' = 180°$. Similarly, $\angle DEB' + B' = 180°$. This leads us to conclude that the sum of the angles in quadrilateral $DEB'C'$ equals $360°$ which contradicts Theorem 17. Thus, our assumption that that $AB < A'B'$ is false. Similarly, we can show that $AB > A'B'$ is false. So, it must be that $AB = A'B'$. In conclusion, triangles ABC and $A'B'C'$ are congruent by the ASA congruence principle (i.e., Book I, Prop. XXVI of Euclid's Elements). ■

. . .

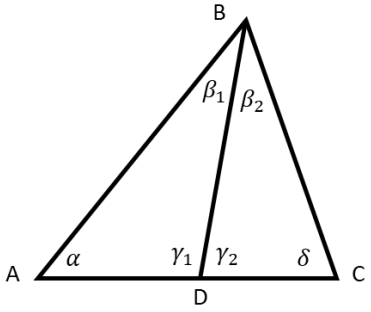
The **deficit of a triangle** is the amount by which the sum of the triangle's angles is less than $180°$. Using this terminology, we can restate Theorem 16 as "There are triangles with deficits arbitrarily close to $0°$."

Consider triangle ABC in the figure below. Let d be the deficit of triangle ABC. Dissect ABC with a line segment from B to D. Let d_1 be the deficit of triangle ABD, and d_2 be the deficit of triangle DBC. By the definition of triangle deficit, we have

$$d = 180° - \alpha - (\beta_1 + \beta_2) - \delta$$
$$d_1 = 180° - \alpha - \beta_1 - \gamma_1$$
$$d_2 = 180° - \delta - \beta_2 - \gamma_2$$

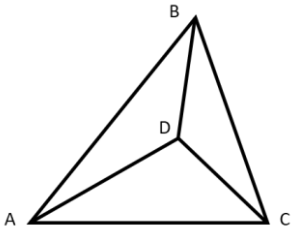
Adding the previous two equations and noting that $\gamma_1 + \gamma_2 = 180°$, we get

$$d_1 + d_2 = 360° - (\alpha + \beta_1 + \beta_2 + \delta) - (\gamma_1 + \gamma_2) = 180° - \alpha - (\beta_1 + \beta_2) - \delta = d$$

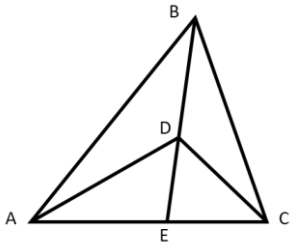


So, for this type of dissection (i.e., line from a vertex of a triangle to the opposite side, known as a transversal), the sum of the deficits of the two resulting triangles equals the deficit of the original triangle.

The same additive property for the deficits of triangles holds true if we divide the triangle with cuts that go from each edge to some interior point, e.g., see the figure below.



For the example at hand, we extend line segment BD until it intersects line AC (see the figure below). In what follows, we use the short hand $\text{def}\,(ABC)$ for the deficit of triangle ABC.



From our previous analysis concerning transversals, we have that

$$\text{def}(ABC) = \text{def}(ABE) + \text{def}(CBE)$$

Iterating on the two smaller triangles, we have

$$\text{def}(ABE) = \text{def}(ADE) + \text{def}\,(ADB)$$

$$\text{def}(CBE) = \text{def}(CED) + \text{def}(CBD)$$

Noting the $\text{def}(ADC) = \text{def}(ADE) + \text{def}\,(CED)$, and using the above results, we have

$$\text{def}(ABC) = \text{def}(ABE) + \text{def}(CBE) = \text{def}(ADB) + [\text{def}(ADE) + \text{def}(CED)] + \text{def}(CBD)$$

$$= \text{def}(ADB) + \text{def}(ADC) + \text{def}(CBD)$$

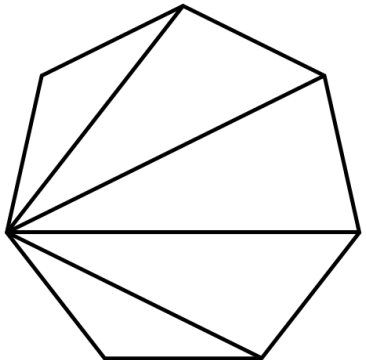
In general, we have the following theorem about triangle deficits.

Theorem 19. If a triangle is subdivided into other triangles in any manner, then the sum of the deficits of the constituent triangles equals the deficit of the given triangle.

In Euclidean geometry, the some of the internal angles of an n-sided polygon is equal to $(n-2) * 180°$ see the Wikipedia article "Polygon" [15]. In hyperbolic geometry, the angle-sum of an n-sided polygon is always less than $(n-2) * 180°$. This follows from Theorem 15 and the fact that a simple polygon (i.e., polygon whose sides do not intersect) can be decomposed into a collection of non-overlapping triangles [16].

Theorem 20. The angle-sum of an n-sided polygon is always less than $(n-2)*180°$.

Proof: Divide the polygon into triangles as follows: Any polygon with n sides (aka an n-gon) can be triangulated by drawing $n-2$ diagonals from one vertex to all the others. This effectively divides the polygon into $n-2$ triangles. For example, the figure below depicts a 7-gon divided into 5 triangles.



Since each triangle has an angle-sum less than $180°$, the sum of the angles for the $n-2$ triangles in the triangulation of the n-gon must be less than $(n-2)*180°$. However, the sum of the angles in the triangles equals the sum of the angles in the n-gon, and so, the sum of the angles in the n-gon must be less than $(n-2)*180°$. ■

. . .

In what follows, we explore the relationship between the summit of a Saccheri quadrilateral and the side of a triangle. Consider side BC of the obtuse triangle ABC in Figure 9 (obtuse angle at vertex B). Let E be the midpoint of side AB, and G be the midpoint of side AC. Draw line EG and then draw perpendiculars from A, B and C to line EG, intersecting line EG at points D, F and H, respectively. By Book I, Prop. XXVI of Euclid's Elements (i.e., the angle-angle-side triangle congruence principle), triangle ADE is congruent to BFE, and triangle ADG is congruent to CHG. Thus, $AD = FB$, and $AD = HC$, which implies $FB = HC$. By definition, $BCHF$ is a Saccheri quadrilateral (upside down) with base FH and summit BC.

The same analysis can be applied to the other two sides of triangle ABC.

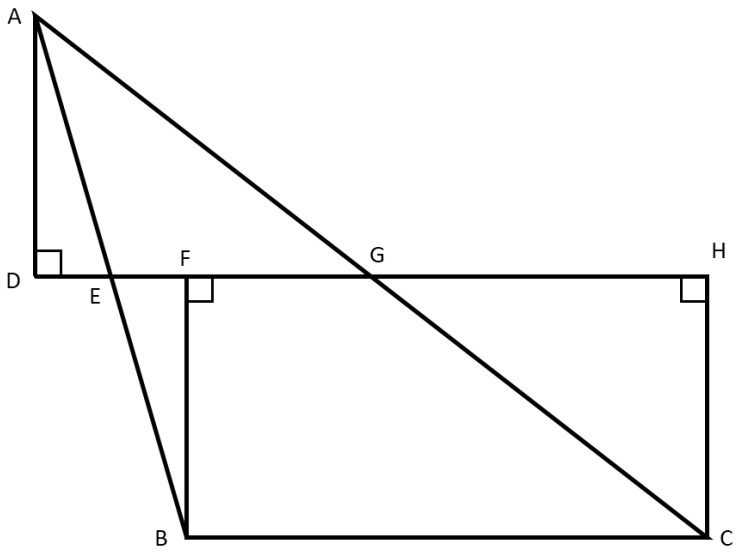


Figure 9. Relationship between side of a obtuse triangle and Saccheri quadrilateral

The above analysis can be applied if triangle ABC has all acute angles (as in Figure 10) as well as for the case where ABC is a right triangle. (The dotted line from B to G is not relevant in the current discussion but will be needed for the discussion in the following section.)

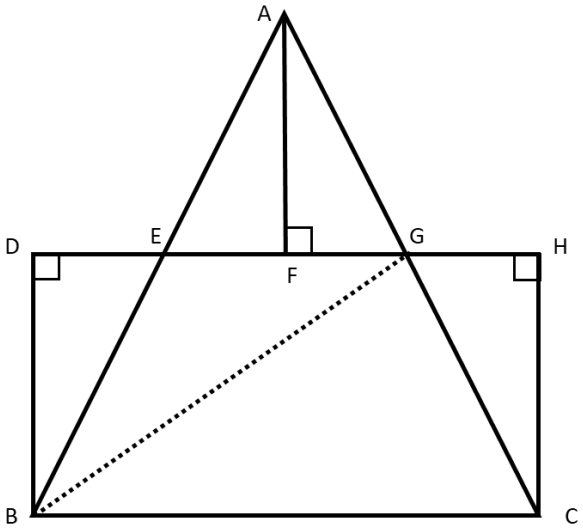


Figure 10. Relationship between sides of an acute triangle and a Saccheri quadrilateral

From the above analysis, the following theorem follows.

Theorem 21. The sum of the angles of a triangle equals the sum of the summit angles in each associated Saccheri quadrilateral.

For example, consider Figure 9. The sum of the summit angles of Saccheri quadrilateral $BCHF$ is $\angle FBC + \angle HCB$. Using the triangle congruences that we established earlier, we have

$$\angle HCB = \angle GCB + \angle HCG = \angle GCB + \angle GAD$$

$$= \angle GCB + \angle DAE + \angle GAE$$
$$= \angle GCB + \angle EBF + \angle GAE$$

So,

$$\angle FBC + \angle HCB = (\angle FBC + \angle EBF) + \angle GCB + \angle GAE$$
$$= \angle EBC + \angle GCB + \angle GAE = \angle B + \angle C + \angle A$$

Using the previous results on Saccheri quadrilaterals, we have the following theorem.

Theorem 22. The line joining the midpoints of two sides of a triangle is parallel to the line containing the third side, has a common perpendicular with it, and the segment joining the midpoints is less than half of the third side.

Proof: The line joining the midpoints of two sides of a triangle is parallel to the line containing the third side since the summit and base of the corresponding Saccheri quadrilateral are parallel by Theorem 2.

The line joining the midpoints of two sides of a triangle has a common perpendicular with the third side of the triangle, since the summit and base of the corresponding Saccheri quadrilateral have a common perpendicular by Theorem 2.

The line joining the midpoints of two sides of a triangle is equal to half the base of the corresponding Saccheri quadrilateral. For example, consider the case where the triangle has an obtuse angle (as shown in Figure 9). We have that

$$FH = DG + GH - (DE + EF)$$

Since $DG = GH$ and $DE = EF$, the above equation reduces to

$$FH = 2(DG - DE) = 2\,EG$$

The calculation is similar for the case where the triangle has acute angles (as shown in Figure 10) and the arguments are even simpler when we have a right triangle.

The segment joining the midpoints of two sides of a triangle is less than half the measure of the third side of the triangle, since and the measure of the base of a Saccheri quadrilateral less than the measure of the summit by Theorem 7. Note that the summit is also the third side of the triangle. ■

2.3.4 Equivalence

Consider Figure 10 from the previous section. Triangle BED and AEF are similar triangles, and CGH and AGF are similar triangles. Further, line segment BG dissect quadrilateral $BEGC$ into the triangles BEG and BCG. So, we have dissected triangle ABC into four triangles that exactly cover Saccheri quadrilateral $BDHC$. The two entities are said to be equivalent. In general, we have the following definition of equivalence in hyperbolic geometry.

> If two polygons can be partitioned into the same finite number of triangles and a one-to-one correspondence can be established so that pairs of corresponding triangles are congruent, the two polygons are said to be **equivalent**.

Using the above definition and the analysis above, we have the following theorem.

Theorem 23. A triangle is equivalent to each of its associated Saccheri quadrilaterals.

The following basic theorem concerns equivalent polygons.

Theorem 24. If two polygons are each equivalent to a third polygon, then they are equivalent to one another.

Proof: See Theorem 1 in Section 62 in Wolfe [10]. ■

The following theorem holds for hyperbolic geometry but not for Euclidean geometry. For example, in Euclidean geometry, similar triangles of different areas are clearly not equivalent.

Theorem 25. Two triangles are equivalent if and only if they have the same angle-sum (or equivalently, the same deficit).

Proof: See Theorems 52 and 53 in Chapter II, Section 11 of Gans [11]. ■

. . .

It is also possible to derive trigonometric formulas in hyperbolic geometry. The interested reader is referred to the Wikipedia article "Hyperbolic triangle" [18], and Chapter VI, Sections 5-7 in Gans [11].

2.3.5 Areas

In Euclidean geometry, if we dissect a **measurable planar set** (i.e., a set of points with a defined area) into several measurable planar sets, then the area of the constituent sets adds to the area of the original set. For example, if we cut a square into two along one of its diagonals, then the sum of the areas of the two resulting triangles equals the area of the square. We would like to preserve this property in hyperbolic geometry.

In the case of triangles, the deficit would give us the additive property described in the previous paragraph. If triangle ABC is divided into triangles ABD and BDC, we know from Theorem 19 that

$$\text{def}(ABC) = \text{def}(ABD \oplus BDC) = \text{def}(ABD) + \text{def}\,(BDC)$$

where $\oplus$ is an operation that composes (joins) two triangles sharing an edge. In fact, we will define the area of a triangle ABC as $Area(ABC) = c \cdot \text{def}(ABC)$ where c is a constant. We still have the additive property if we use a constant in the definition, i.e.,

$$\begin{aligned} Area(ABC) &= c \cdot \text{def}(ABC) = c \cdot [\text{def}(ABD \oplus BDC)] \\ &= c \cdot \text{def}(ABD) + c \cdot \text{def}(BDC) \\ &= Area(ABD) + Area(BDC) \end{aligned}$$

In general, if $f(x + y) = f(x) + f(y)$ and f is a continuous function, then the only solution is of the form $f(x) = cx$ with c being a constant. This is known as Cauchy's functional equation [17]. As is, our problem does not actually fit Cauchy's functional equation. However, by Theorem 25, we know that triangles with the same angle-sum are equivalent. So, we can replace a triangle by its angle-sum x in our area formula property, i.e., $Area(x + y) = Area(x) + Area(y)$. Viewed in this way, we can apply the result concerning Cauchy's functional equation.

Based on our definition of area, two triangles have the same area if and only if they have the same defect, and by Theorem 25, if and only if they have the same angle-sum. We record this fact in the following theorem.

Theorem 26. Two triangles have the same area if and only if they have the same angle-sum.

Clearly, this is much different from the case in Euclidean geometry where all triangles have the same angle-sum (i.e., $180°$) but can have different areas. If we take (for example) $c = 1$, then the largest possible area for a triangle in hyperbolic geometry would be less than $180°$ or π radians (which corresponds to an angle-sum approaching 0).

. . .

We can extend the definition of area to polygons as follows:

> The area of a polygon is the sum of the areas of all of the triangles of any partition of the polygon into triangles.

If the difference between $(n-2) * 180°$ and the sum of the angles of a polygon of n sides is defined to be the **defect of the polygon**, then it follows that the defect of a polygon is equal to the sum of the defects of all of the triangles in any partition of the polygon. The following theorems, concerning the area of a polygon, all rely on the fact that a polygon can be decomposed into a collection of non-overlapping triangles.

Theorem 27. Two polygons have the same area if and only if they are equivalent.

Theorem 28. If a polygon is partitioned into triangles in any way, the area of the polygon is equal to the sum of the area of all of the triangles in the partition.

Theorem 29. If a polygon is equivalent to two or more component polygons, the area of the polygon is equal to the sum of areas of the component polygons.

2.3.6 Circles

In hyperbolic geometry, a circle is defined as the set of all points that are the same distance (i.e., the length of the radius of the circle) from a given point (i.e., the center of the circle). Other terms and concepts from Euclidean geometry also apply to circles in hyperbolic geometry, e.g., diameter, arc, chord, secant, tangent, central angle, inscribed angle, congruent circles. To refresh your memory of these terms, see Book III. Theory Of The Circle in Euclid's Elements [12].

All the theorems about circles from Euclidean geometry whose proofs do not rely on Euclid's Postulate #5 still holds true in hyperbolic geometry. The following are some of the theorems (paraphrased from Book III of Euclid's Elements) which hold true for circles in hyperbolic geometry.

- (Prop. II.) Given any two points on the circumference of a circle, all the points on the chord determined by the two points must fall within the circle.
- (Prop. III.) If a line passing through the center of a circle bisects a chord (which is not a diameter of the circle), then the line intersects the chord at a right angle. Conversely, if a line passing through the center of a circle intersects a chord (which is not a diameter of the circle) at a right angle, then the line bisects the chord.
- (Prop. IV.) Two chords of a circle which are not both diameters cannot bisect each other, though either may bisect the other.
- (Prop. XI. and XII.) Given two tangent circles, their centers and the point of intersection lie on a line. This holds whether one circle is internal to the other or if the circles are external to each other.

- (Prop. XVIII.) At the point P of tangency of a line ℓ to a circle, a radius drawn to point P is perpendicular to ℓ.
- (Prop. XIX.) If line ℓ is tangent to a circle at point P, the line m perpendicular to ℓ at point P must pass through the center of the circle.
- (Prop. XXIX) In the same circle or congruent circles, equal central angles subtend equal chords.

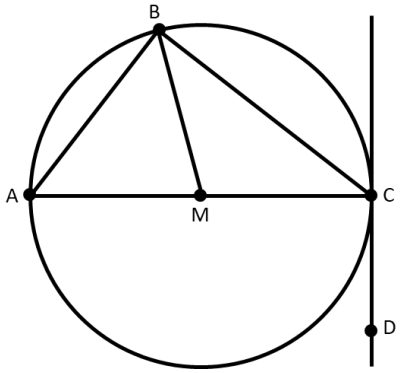
The following are some properties of circles (from Book III of Euclid's Elements) that hold true for Euclidean geometry but not for hyperbolic geometry.

- (Book III, Prop. XXXI.) The angle inscribed in a semicircle is a right angle.
- (Book III, Prop. XX.) An inscribed angle is measured by half its intercepted arc.
    - In hyperbolic geometry, an angle inscribed in a circle is less than half the central angle subtending the same arc
- Any three noncollinear points determine a unique circle.

In contrast to Prop. XXXI (noted above), we have the following theorem in hyperbolic geometry.

Theorem 30. The angle inscribed in a semicircle is acute and varies depending on its location.

Proof: In the figure below, we have a circle with diameter AC and center at point M. Let B be a point on the circle other than A or B. By definition, ABC is inscribed in the semicircle comprising the upper half of the circle in the figure.



By Theorem 15, $\angle A + \angle B + \angle C < 180°$. Triangle ABM is isosceles since it has two equal sides (i.e., AM and BM are radii of the circle), and so, $\angle A = \angle ABM$. Similarly, BMC is isosceles and so, $\angle C = \angle MBC$. Thus,

$$\angle A + \angle C = \angle ABM + \angle MBC = \angle B$$

Substituting the above into the inequality $\angle A + \angle B + \angle C < 180°$ gives us $\angle B < 90$, which proves the first part of the theorem.

Concerning the second part of the theorem, let line CD be tangent to the circle at point C. Keep points A and C fixed in position, and vary the location of point B along the semicircle. As B approaches C, lines AB and BC approach lines AC and CD, respectively. Further, $\angle B$ approaches $\angle ACD = 90°$ as B approaches C. So, if $\angle B$ had a constant value for all locations of B on the upper

semicircle, the value would have to be $90°$ but this contradicts the first part of the theorem. Thus, the measure of $\angle B$ must vary. ■

2.3.7 Parallel Lines without a Common Perpendicular

In Section 2.3.2, we studied parallel lines that shared a common perpendicular. In this section, we'll discuss the case of parallel lines that do not share a common perpendicular. As we shall show, given a line ℓ and point P not on ℓ, there are exactly two lines through P that are parallel ℓ but do not share a common perpendicular with ℓ.

Given a line ℓ and a point P not on line, we know from Theorem 13 that there are an infinite number of lines through P that are parallel to ℓ and have a common perpendicular with ℓ. In Figure 11, lines h, g, n pass through point P and are parallel to line ℓ. There are also lines that pass through P that do intersect ℓ (e.g., line PB in the figure). [**Author's Remark**: the lines are draw slightly curved to help the reader visualize how, for example, line n does not intersect line ℓ.]

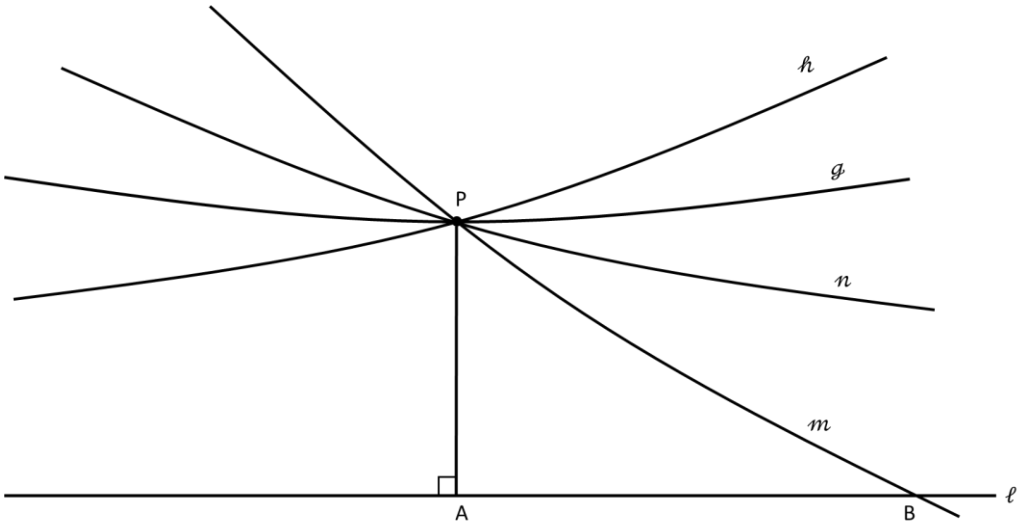


Figure 11. Parallel lines to a given line and through a given point

Regarding Figure 12, if A is the projection of P onto line ℓ, and g is the line through P at a right angle to line AP, then there is a set of infinitely many lines that subdivide $\angle APD$ and that have a common perpendicular with ℓ (this follows from the construction in the proof of Theorem 13 but not explicitly stated in the theorem). Similarly, there is a set of infinitely many lines that subdivide $\angle APC$ and that have a common perpendicular with ℓ. In each of these sets of lines, there is no line that is closest to line AP in the sense of making a smaller angle with it than do all the others (follows from Theorem 14).

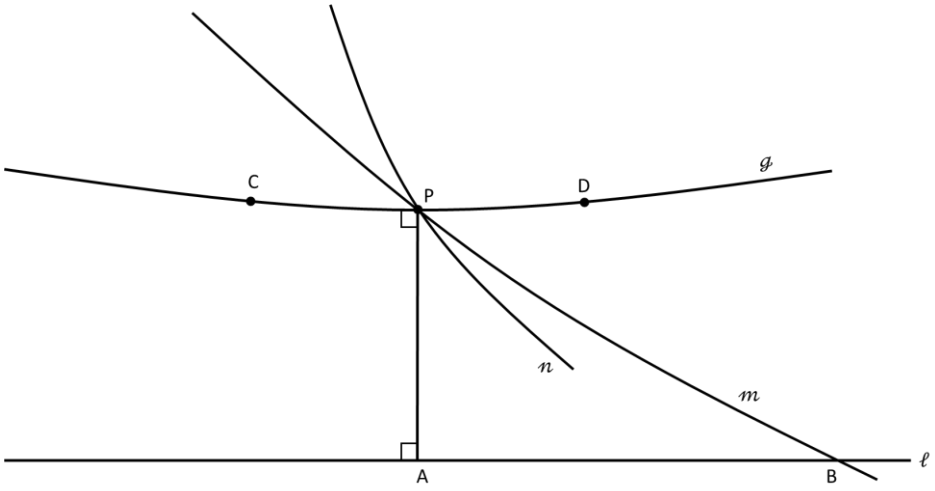


Figure 12. Subdividers of right angle

Next, consider the set S of all the lines which subdivide $\angle APD$. If we take two distinct members from S, then one of the two will make a smaller acute angle with line AP. The line making the smaller acute angle with AP is said to *precede* the other. For example, in Figure 12, line n precedes line m. Set S can be split into two distinct subsets, i.e.,

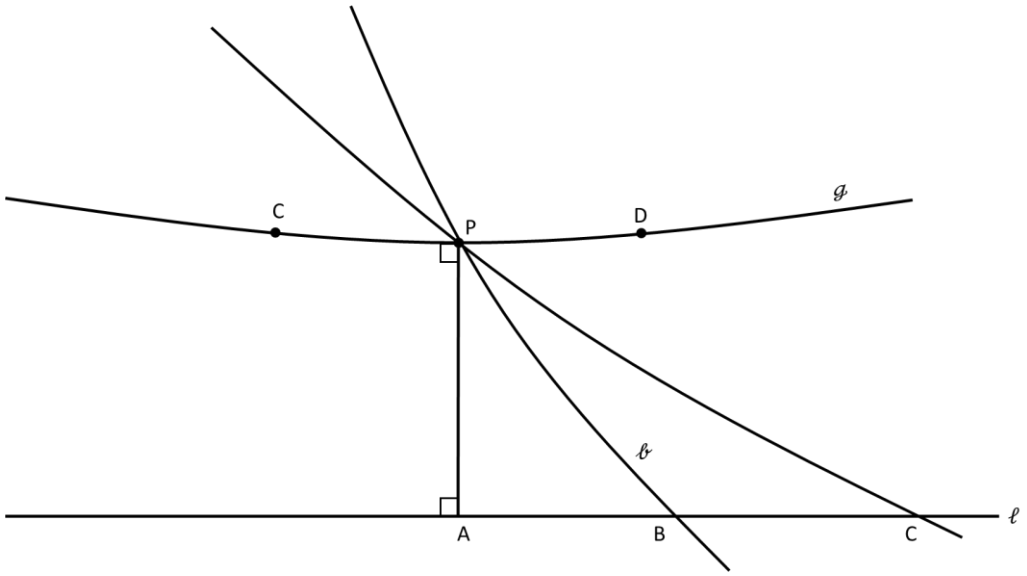
- Set M: the subdividers which meet (i.e., intersect) line ℓ
- Set N: the subdividers that do not meet line ℓ.

Every member of M precedes every member of N. (To see this, assume the statement is false and that some line $n \in N$ does precede some line $m \in M$, see Figure 12. In this case, n makes a smaller angle with line AP than does m. Since $m \in M$, there exist a point B where m meets ℓ. Hence, n subdivides $\angle APB$ of triangle APB and thus, intersects side AB. This contradicts the assumption that $n \in N$.)

In general, the following property holds true in both Euclidean and hyperbolic geometry.

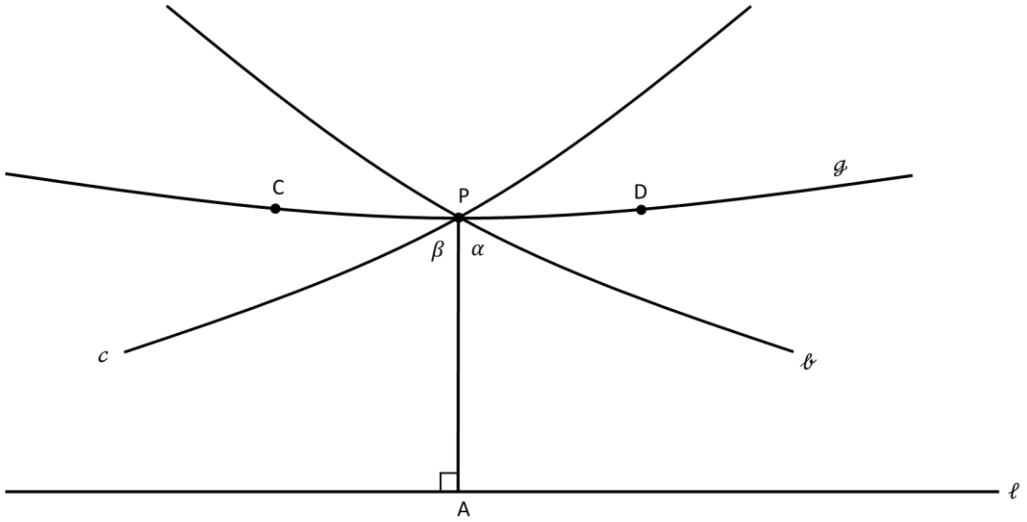
> If the lines subdividing an angle are formed into two sets A and B such that each line in A precedes each line in B, then either A has a line preceded by every other line in A (a "last line" so to speak), or B has a line preceding every other line in B (a first line). This particular subdivider is called the boundary between A and B.

From the above property, there is a subdivider $\mathscr{b}$ which is the boundary between set M and N. The boundary subdivider $\mathscr{b}$ must be in N. To prove this, assume to the contrary that $\mathscr{b} \in M$. In this case, $\mathscr{b}$ meets ℓ in some point B as shown in the figure below. Next, select any point C on line ℓ such that B is between A and C. By definition, the line PC is in M, and is preceded by $\mathscr{b}$ since $\mathscr{b}$ makes a smaller angle with line AP than does line PC. However, $\mathscr{b}$ being the boundary subdivider it must be preceded every other line in M and so, we have a contradiction, and must conclude that $\mathscr{b} \in N$.

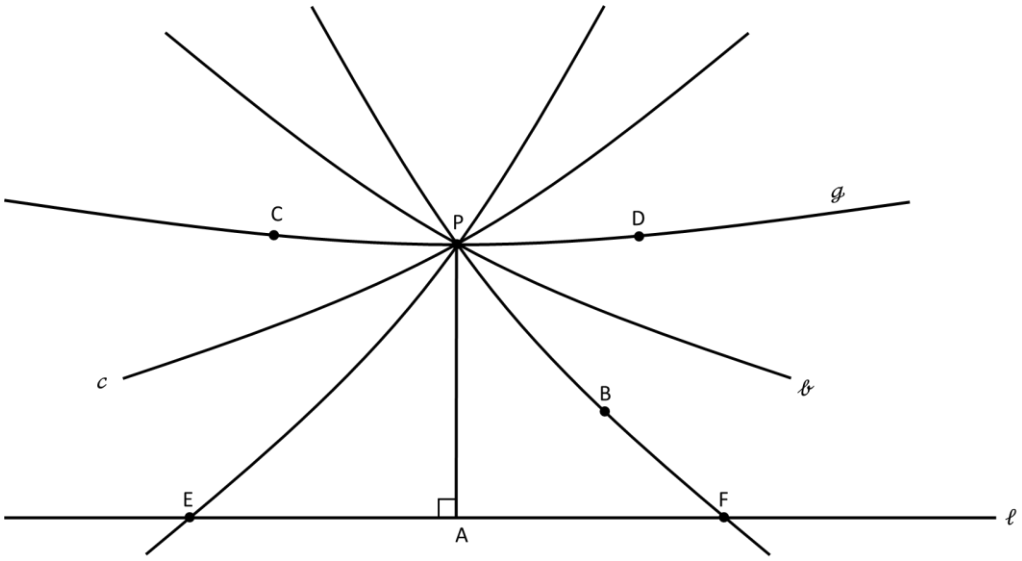


So, $\mathscr{b}$ is a line which does not meet ℓ and precedes all the other lines in set N and thus makes a smaller angle with line AP than all other lines through P that are parallel to ℓ. By Theorem 14, $\mathscr{b}$ cannot have a common with ℓ.

A similar discussion to that above, as applied to the left side of the previous figures, would show there is another boundary line c, which is parallel to ℓ, and has no common perpendicular with ℓ. The two boundary lines, along with the angles they make with line AP (i.e, angles α and β), are shown in the figure below.

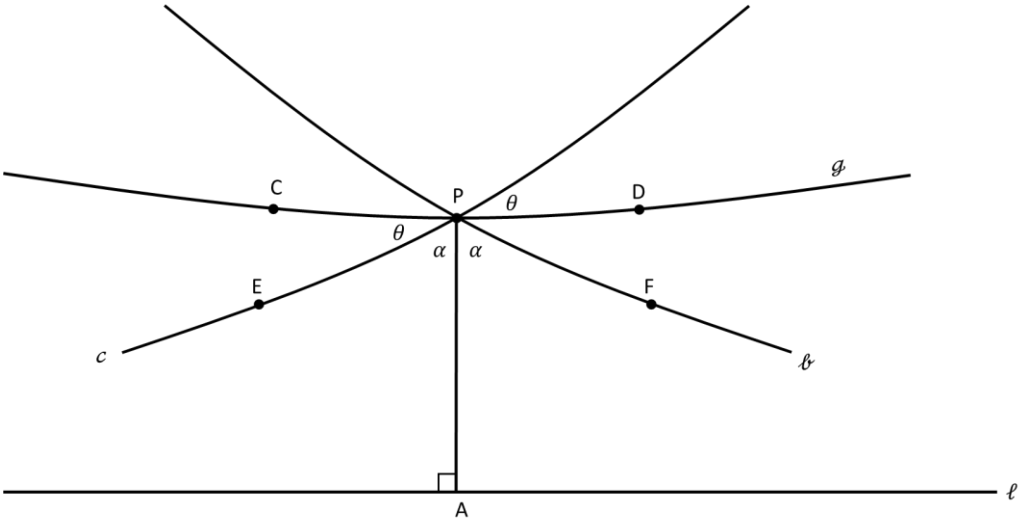


Given the symmetry of the arguments for the left and right sides of the previous figures, one would expect that $\alpha = \beta$. To prove this, let line PB subdivide α such that $\angle APB = \beta$ (see the figure below). Noting that line PB must be in set M, let F be the point of intersection of PB with ℓ. Select E on ℓ such that A is the midpoint of EF. By the SAS triangle congruence principle, triangles AEP and AFP are congruent. Thus, $\angle FPA = \angle EPA = \beta$ which implies that PE coincides with boundary line c and thus, we have a contradiction since we know that line c does not intersection line ℓ. Thus, our initial assumption is false, and it must be that $\alpha = \beta$.



In summary, we have the following theorem.

Theorem 31. Given line ℓ and any point P not on it, there exist exactly two lines b and c which go through P, are parallel to ℓ, but do not have a common perpendicular with ℓ. If A is the projection of P onto ℓ, then b and c make equal acute angles with line AP. Within one pair of equal vertical angles (the angles marked α in the figure below) lie all the lines through P which meet ℓ, and within the other pair of vertical angles (the angles marked θ in the figure below) lie all the parallels to ℓ through P which have a common perpendicular with ℓ.

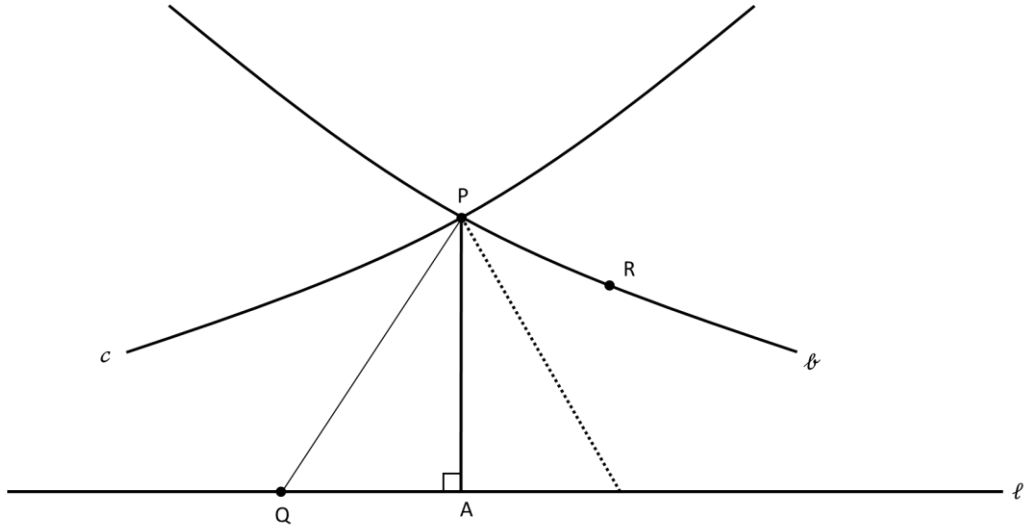


The lines b and c in the above figure will be referred to as the **boundary parallels** to ℓ through point P. The other parallels to to ℓ through point P are called **non-boundary parallels**. The equal angles α (on each side of line AP) are called the **angles of parallelism** corresponding to line ℓ and point P. The sides of these angles which lie on b and c are called **boundary rays**. In the figure above, $\overrightarrow{PF}$ and $\overrightarrow{PE}$ are boundary rays (arrow added to emphasize direction).

We have the following theorem concerning boundary rays.

Theorem 32. Consider the boundary parallels $\mathscr{b}$ and c to line ℓ through point P. If Q is any point on ℓ, and R is any point on a boundary ray, then every line subdividing $\angle QPR$ intersects ℓ.

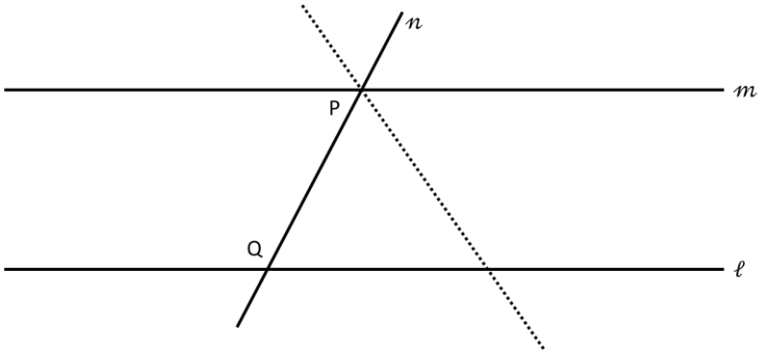
The figure below shows an example of the property stated in Theorem 32. The dotted line subdivides $\angle QPR$ and also intersects line ℓ.



The following theorem provides a condition for when a line is a boundary parallel to another line.

Theorem 33. If two non-intersecting lines m and ℓ are intersected by a transversal n in points P and Q, respectively, and every line subdividing one of the interior angles at P intersects m, then m is a boundary parallel to ℓ going through point P.

An example of Theorem 33 is shown in the figure below. The dotted line is a subdivider of what is referred to as an interior angle at P.



We can speak of boundary parallels to a given line as having one of two directions. In Figure 13, there are multiple boundary parallels to line ℓ that go through various points. Boundary parallels $\mathscr{b}, e$ and $\mathscr{g}$ all point in the same direction in the sense that each makes an acute angle to the right of the perpendicular dropped from points P, Q and R to line ℓ. Boundary parallels $c, \mathscr{d}$ and $\mathscr{f}$ all point in the same direction in the sense that each makes an acute angle to the left of the perpendicular dropped from points P, Q and R to line ℓ. We could, for example, rotate the entirety

of Figure 13 $90°$ clockwise and still talk about two directions (up and down in this case). The point is that there are two opposite directions for the boundary parallels relative to a given line – these are known as **directions of parallelism**.

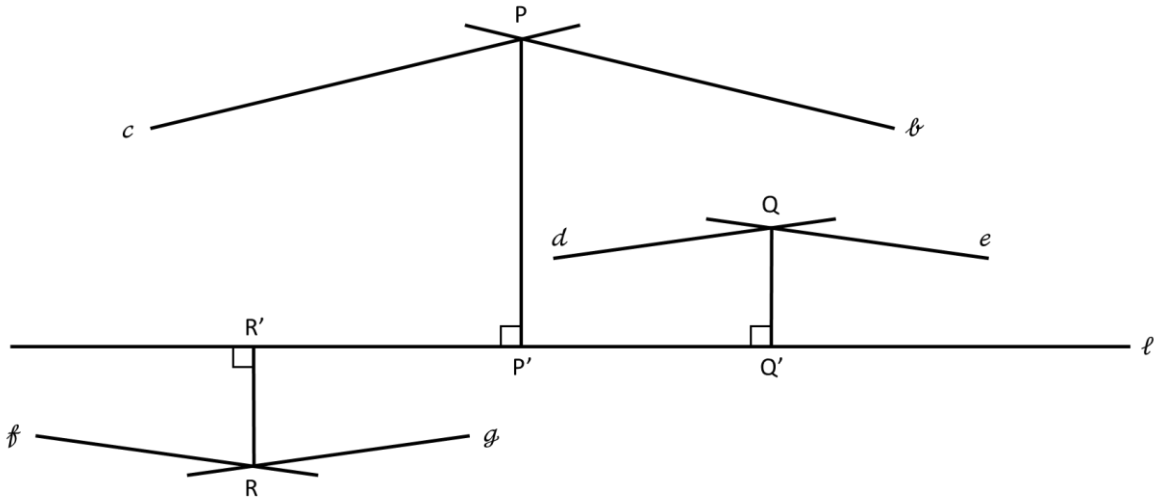


Figure 13. Direction of boundary parallels

The following are three basic theorems regarding parallel lines and boundary parallels. We state them without proof. The interested reader can find proofs in Chapter IV, Section 3 of Gans [11] and in Section 38 of Wolfe [10].

Theorem 34. If a line is the parallel, through one of its points, to another line in a given direction, then it is the parallel, through each of its points, to that line in that same direction.

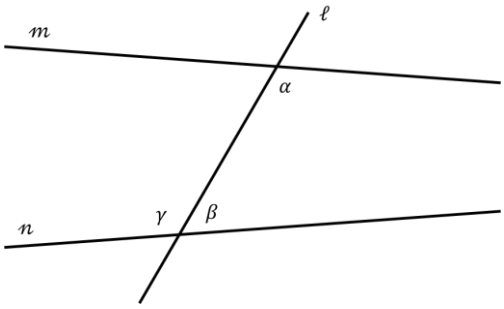
Theorem 35. If line m is a boundary parallel to line n, then line n is a boundary parallel to line m. Further, the boundary rays on each line project onto the boundary rays of the other line, i.e., both lines have the same direction of parallelism.

Theorem 36. If two lines are boundary parallels to a third in a given direction, then they are boundary parallels to each other, with the direction of parallelism being the same for all three lines.

Given Theorem 35, we can speak of two boundary parallels relative to each other in lieu of saying "two boundary parallels to a given line". With this caveat in mind, we state the following theorem.

Theorem 37. If two boundary parallels (relative to each other) are intersected by a transversal, and α and β are the interior angles such that one side of each is a boundary ray, then $\alpha + \beta < 180°$.

Proof: The situation described in the theorem is shown in the figure below. By Theorem 35, the boundary rays of the two lines have the same direction.

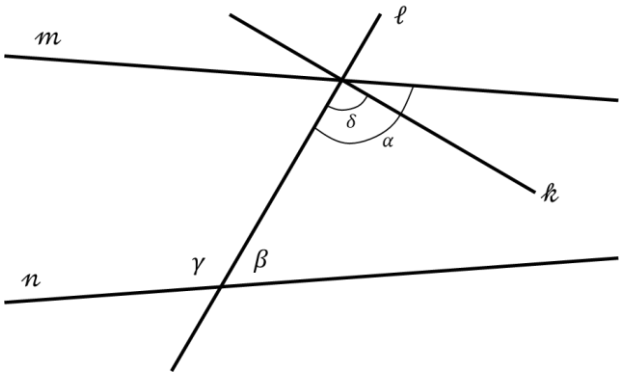


There are several cases.

If the transversal ℓ is perpendicular to one of these parallels (i.e., line m or n in the above figure), then one of the interior angles is $90°$ and the other is an angle of parallelism (which would be acute by Theorem 31). Thus, the sum of the two angles is less than 180°.

Next, suppose the transversal ℓ is not perpendicular to m or n.

- Assume $\alpha + \beta = 180°$. We also know that $\beta + \gamma = 180°$ (straight line). Thus, the alternate interior angles α and γ are equal, and the lines m and n have a common perpendicular (by Theorem 10), which contradicts the assumption that m and n are boundary parallels.
- Assume $\alpha + \beta > 180°$. Place line k such that it subdivides α, and $\delta + \beta = 180°$. Thus, $\gamma = \delta$, and n and k are parallels with a common perpendicular (by Theorem 10). However, this is impossible since k is a subdivider of α, and must intersect n (by Theorem 32).



So, we are left with the only possibility, i.e., $\alpha + \beta < 180°$. ■

2.3.8 Trilaterals

If we consider two boundary rays each associated with a boundary parallel in the same direction, and a transversal segment, we get something called a **trilateral**, see the Figure 14. In the figure, boundary rays AC and BD are referred to as the outer sides of the trilateral. Line segment AB is referred to as the inner (or middle side) of the trilateral. The angles $\angle BAC$ and $\angle CAB$ are the interior angles of the trilateral. The trilateral also has two exterior angles, one of which is shown in the figure, i.e., angle γ. Points A and B are the vertices of the trilateral. The trilateral in the figure is referred to as "trilateral $CABD$".

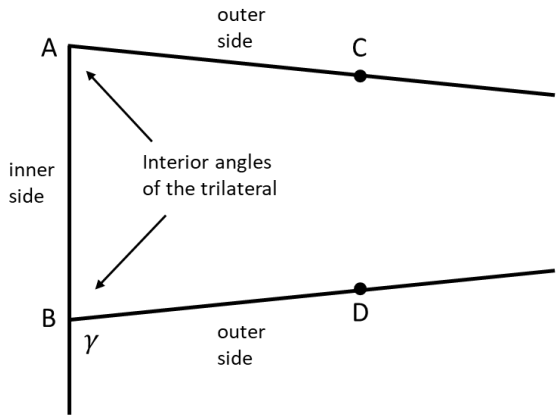


Figure 14. Trilateral and associated terminology

Trilaterals have similar properties to triangles in hyperbolic geometry, e.g.,

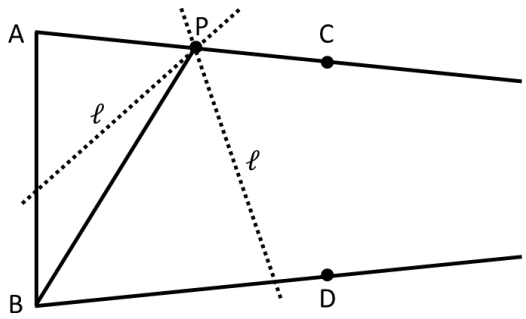
- Trilaterals have three sides.
- The angles of a trilateral add to less than $180°$ (by Theorem 37).
- A line which subdivides an angle of a trilateral meets the opposite side (by Theorem 32).

For triangles, if a line intersects one side and does not intersect any vertices of the triangle, the line must intersect another side. With an additional condition, the same is true for trilaterals.

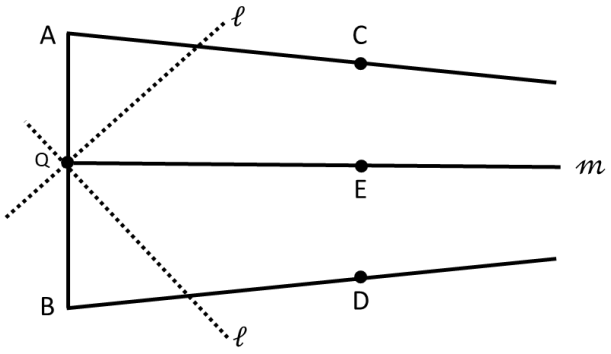
Theorem 38. A line which intersects one side of a trilateral, and does not intersect any vertices, will intersect another side of the trilateral, provided that the line is not a boundary parallel to an outer side.

Proof: There are two cases, i.e., the line intersects an outer side, or the inner side.

Assume the line (call it ℓ) intersects the outer side AC at point P (see the figure below). Line ℓ subdivides $\angle APB$ or $\angle BPC$. If ℓ subdivides $\angle APB$, then it intersects side AB of triangle APB (basic property of triangles). If ℓ subdivides $\angle BPC$, then it must intersect boundary parallel BD (by Theorem 32).



Assume the line ℓ intersects the inner side AB at point Q (see the figure below). Let m be the boundary parallel to line AC through point Q. By Theorem 36, m is also a boundary parallel to line BD.



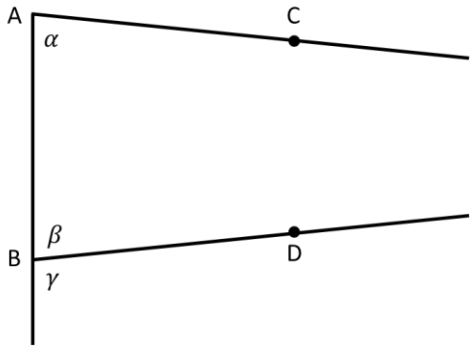
By hypothesis, line ℓ is not a boundary parallel to either outer side of the trilateral, and thus distinct from line m. So, ℓ either subdivides $\angle AQE$ or $\angle BQE$. If ℓ subdivides $\angle AQE$, then by Theorem 32, it intersects line AC. If ℓ subdivides $\angle BQE$, then by Theorem 32, it intersects line BD.

So, in all cases, line ℓ intersects another side of the trilateral. ■

The following theorem concerns the exterior angle of a trilateral. The result is similar to that for triangles.

Theorem 39. In a trilateral, an exterior angle is greater than the opposite interior angle.

Proof: Consider the exterior angle γ in the trilateral shown in the figure below.



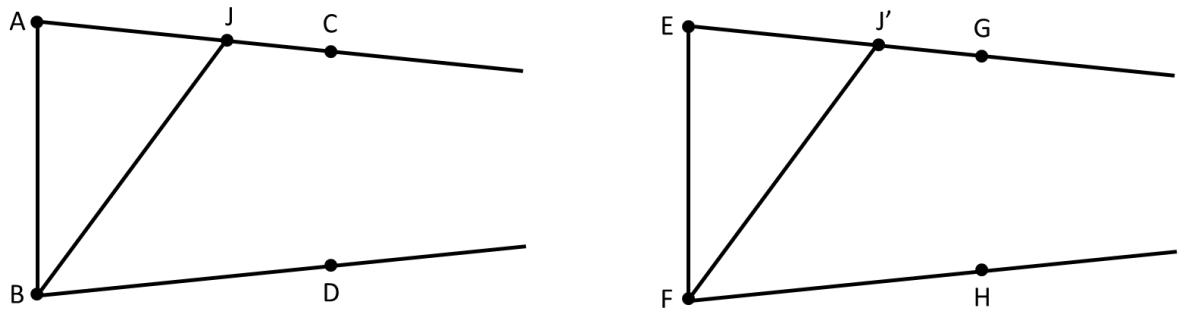
Since AB is a straight line, we know that $\beta + \gamma = 180°$. We also know that the sum of the angles of a trilateral is less than $180°$, i.e., $\alpha + \beta < 180°$. So, $\beta + \gamma > \alpha + \beta$ which implies $\gamma > \alpha$. ■

We can also define **congruence for trilaterals**, i.e., two trilaterals are congruent if the angles and middle side of one are equal in measure to the angles and middle side of the other. This is the Side-Angle-Side (SAS) congruence principle for trilaterals.

The equivalents of the following two theorems are not necessarily true for triangles.

Theorem 40. If an angle and the middle side of one trilateral are equal to an angle and the middle side of another trilateral, then the two trilaterals are congruent.

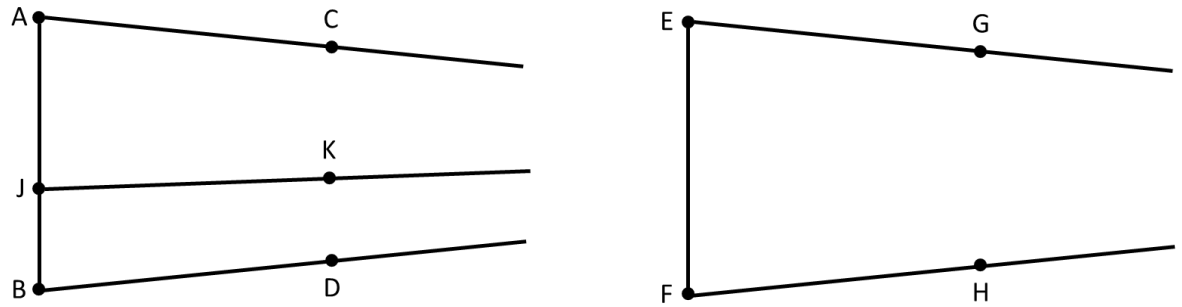
Proof: In the figure below, assume that trilaterals $CABD$ and $GEFH$ are such that $AB = EF$ and $\angle A = \angle E$.



Assume $\angle B > \angle F$. Choose the point J on AC such that $\angle ABJ = \angle F$, and choose J' on line EG such that $AJ = EJ'$. By the SAS triangle congruence principle, triangles ABJ and EFJ' are congruent which implies that $\angle ABJ = \angle EFJ'$. However, $\angle ABJ = \angle F$ and so, $\angle F = \angle EFJ'$ but this implies that FJ' coincides with FH which is impossible. So, our assumption that $\angle B > \angle F$ must be false. By similar arguments, the assumption that $\angle B < \angle F$ is also false. So, we must conclude that $\angle B = \angle F$, and thus, trilaterals $CABD$ and $GEFH$ are congruent since their corresponding interior angles and middle side are equal to each other. ■

Theorem 41. If the corresponding interior angles of one trilateral are equal to those of another, then the two trilaterals are congruent.

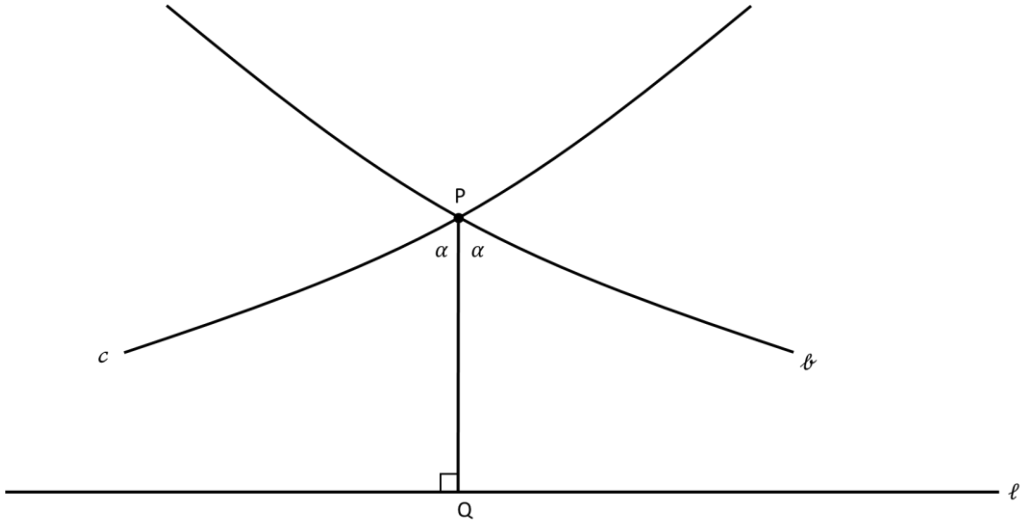
Proof: Let $CABD$ and $GEFH$ be trilaterals with equal interior angles, i.e., $\angle A = \angle E$ and $\angle B = \angle F$. To prove the two trilaterals are congruent, we need to show that $AB = EF$. Assume to the contrary, i.e., $AB > EF$.



Select point J on AB such that $AJ = EF$. Let line JK be the boundary parallel to line BD through point J. By Theorem 40, trilateral $CAJK$ is congruent to trilateral $GEFH$ which implies $\angle AJK = \angle EFH$. Further, $\angle AJK$ is an exterior angle to trilateral $KJBD$ but $\angle AJK = \angle EFH = \angle ABD$. So, we have an exterior angle to a trilateral equal to the opposite interior angle of the same trilateral, which contradicts Theorem 39. Thus, our assumption that $AB > EF$ is false. In an analogous manner, we can show that the assumption $AB < EF$ also leads to a contradiction. Thus, it must be that $AB = EF$ and therefore, $CABD$ and $GEFH$ are congruent. ■

2.3.9 Alternate View of Boundary Parallels

As we saw earlier, a given line ℓ and point P on the line determines two equal angles of parallelism. The angles of parallelism are acute as measured relative to a line perpendicular ℓ and through point P (see line PQ in the figure below).



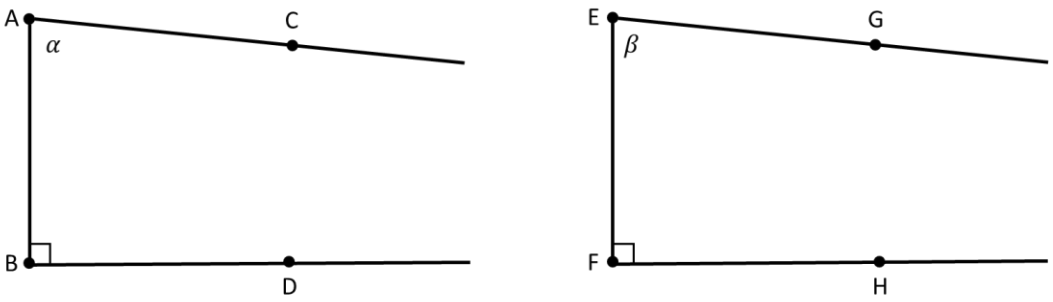
Alternately, we could start with the line segment PQ, draw a line ℓ perpendicular PQ at point Q, and then determine the two boundary parallels to line ℓ through point P. The first part of the following is an alternative statement of Theorem 31 in terms of line segments. The second sentence in the theorem is the converse of the first sentence (which we will prove later).

Theorem 42. For any line segment PQ there corresponds two equal angles of parallelism and they are acute. For every acute angle there is an angle of parallelism corresponding to some segment.

We make use of the trilateral concept to prove the following theorem.

Theorem 43. If two line segments are equal, their corresponding angles of parallelism are equal. Conversely, if two angles of parallelism are equal, so are the segments to which they correspond.

Proof: For the first part of the theorem, we are given two equal line segments, i.e., AB and EF. Draw a line perpendicular to AB at point B, and a line perpendicular to EF at point F. Draw a right boundary parallel to BD going through point A, and a right boundary parallel to FH through point E, as shown in the figure below. By Theorem 40, trilaterals $CABD$ and $GEFH$ are congruent, which implies that $\angle BAC = \angle FEG$, i.e., the two segments have equal angles of parallelism.

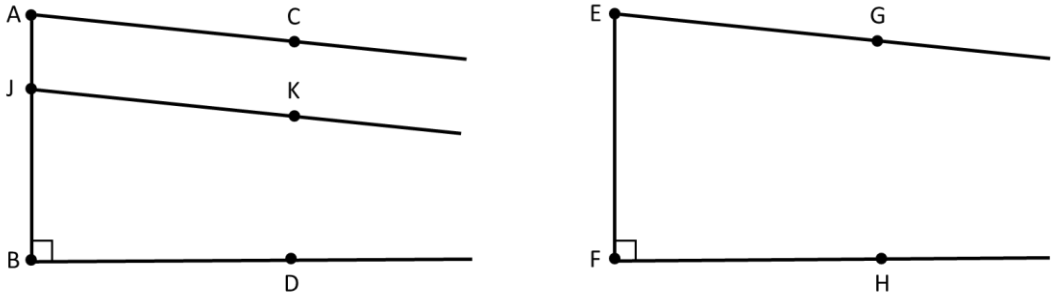


Going in the other direction, assume that we have two equal angles of parallelism, i.e., $\alpha = \beta$. By Theorem 42, there is segment AB that has corresponding angle of parallelism α and a segment EF that has corresponding angle of parallelism β. Draw a perpendicular line to AB at point B, and a perpendicular line to EF at F, as shown in the figure above. By Theorem 41, trilaterals $CABD$ and $GEFH$ are congruent, and thus $AB = EF$. ■

The next theorem states that the size of a line segment is inversely related to the measure of its corresponding angle of parallelism.

Theorem 44. Given line segments AB and EF such that $AB > EF$, the corresponding angle of parallelism for AB is less than the corresponding angle of parallelism for EF, and conversely.

Proof: Consider the trilaterals $CABD$ and $GEFH$ in the figure below, with AB perpendicular to BD and EF perpendicular to FH. Further, $\angle A$ is the angle of parallelism with respect to AB, and $\angle E$ is the angle of parallelism with respect to EF. Select point J on AB such that $JB = EF$. Select K such that JK is parallel to line BD. By Theorem 40, trilateral $KJBD$ is congruent to $GEFH$ which implies that $\angle BJK = \angle FEG$. Since $\angle BJK$ is an exterior angle to trilateral $CAJK$, $\angle BJK > \angle JAC$ by Theorem 39. So, $\angle FEG = \angle BJK > \angle JAC = \angle BAC$ which proves the theorem in the forward direction.



Let α and ε be two angles of parallelism such that $\alpha < \varepsilon$. By Theorem 42, there exists line segment AB corresponding to α, and line segment EF corresponding to ε. Theorem 43 implies that $AB \neq EF$. If $AB < EF$, then the first part of this theorem would imply that $\alpha > \varepsilon$ (which contradicts our initial assumption). Thus, it must be the $AB > EF$. ■

. . .

Shown below is the Lobachevskii function [19]. It takes as input the length of a line segment x and outputs the corresponding angle of parallelism α in radians. The capital Greek letter Pi is used to represent this function.

$$\alpha = \Pi(x) = 2\arctan e^{-\frac{x}{k}}$$

We have that $\Pi(0) = \pi/2$. As x increases the corresponding angle of parallelism $\Pi(x)$ decreases and approach 0 as x approaches infinity. In the equation, $k > 0$ is a constant that determines the fixed scale of measurement. We can also solve for x in terms of the corresponding angle of parallelism, i.e.,

$$\tan\left(\frac{\alpha}{2}\right) = e^{-\frac{x}{k}}$$

$$\ln\left(\tan\left(\frac{\alpha}{2}\right)\right) = -\frac{x}{k}$$

$$x = k\ln\left(\cot\left(\frac{\alpha}{2}\right)\right)$$

So, for every acute angle there is an angle of parallelism corresponding to some segment, i.e., the second part of Theorem 42.

2.3.10 Distance between Lines

In both Euclidean and hyperbolic geometry, the distance between two intersecting lines becomes greater as one moves away from the point of intersection. We state this more formally in the theorem below.

Theorem 45. Given two intersecting lines, the perpendicular distance (i.e., shortest distance) from a point on one of them to the other increases without limit as the point moves away from the point of intersection, and becomes smaller as the point moves toward the intersection.

Proof: See Theorem 18 in Chapter IV, Section 6 of Gans [11], or Theorem 1 in Section 47 of Wolfe [10].

The following theorem applies to parallel lines with a common perpendicular (in hyperbolic geometry).

Theorem 46. Given two parallel lines with a common perpendicular, the shortest distance from a point on one line to the other becomes arbitrarily great as the point recedes from that perpendicular in either direction.

Proof: See Theorem 19 in Chapter IV, Section 6 of Gans [11], or Theorem 3 in Section 47 of Wolfe [10].

On the other hand, if two parallel lines (in hyperbolic geometry) do not have a common perpendicular, they become arbitrarily close in one direction and arbitrarily far apart in the other direction. More formally, we have the following theorem.

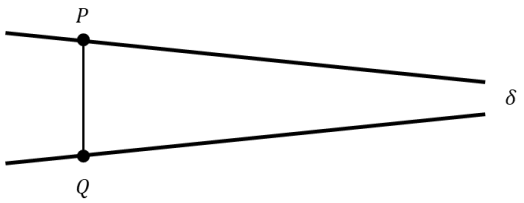
Theorem 47. Two parallel lines (not having a common perpendicular) converge continuously in the direction of parallelism and diverge continuously in the opposite direction.

Proof: See Theorem 20 in Chapter IV, Section 6 of Gans [11], or Theorem 2 in Section 47 of Wolfe [10].

Two boundary parallels are asymptotic to each other (i.e., become arbitrarily close) in their common direction of parallelism. For this reason, boundary parallels are sometimes referred to as **asymptotic parallels**. In the same vein, parallels with a common perpendicular are called non-asymptotic parallels.

Since two boundary parallels are asymptotic in their directions of parallelism, we can think of these directions as determining one direction in the plane, and so, we can speak of the two lines as being parallel in the same direction. In conjunction with this idea, we introduce the notation $P\delta$ to indicate a line going through point P and in the direction δ.

If two boundary parallels go through distinct points P and Q, we can represent the two lines as $P\delta$ and $Q\delta$, where δ represents the common direction in the plane determined by the direction of parallelism of the two lines (see the depiction in the figure below). Further, the trilateral with inner side PQ and outer sides on $P\delta$ and $Q\delta$ can be denoted as $PQ\delta$ or $QP\delta$.



2.3.11 Perpendicular Bisectors of a Triangle

In Euclidean geometry, the perpendicular bisectors of each side of a triangle meet at a point, see Figure 15. The figure below and a proof of this fact can be found in the Proof Wiki article "Perpendicular Bisectors of Triangle Meet at Point" [20].

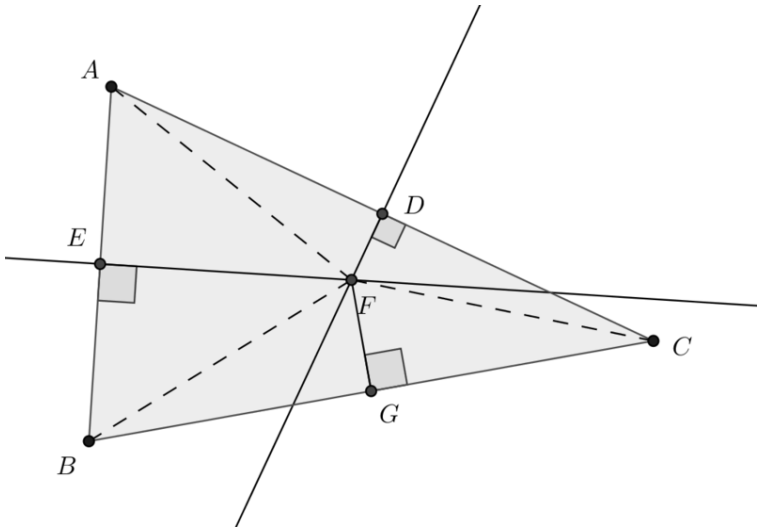


Figure 15. Perpendicular bisectors of a triangle

In hyperbolic geometry, there are three cases concerning the perpendicular bisectors of a triangle. The following theorem summarizes the three cases.

Theorem 48. The perpendicular bisectors of a triangle either intersect at a point, are parallel with a common perpendicular, or are boundary parallels in the same direction.

Proof: See Chapter IV, Section 8 of Gans [11].

An example of when the perpendicular bisectors are parallel with a common perpendicular is shown in Figure 16. The common perpendicular is line ℓ.

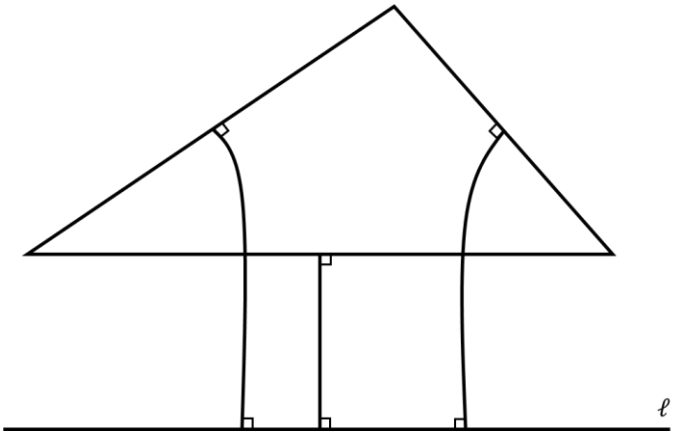


Figure 16. Perpendicular bisectors in parallel and with a common perpendicular

An example of when the perpendicular bisectors are boundary parallels in the same direction is shown in Figure 17. As suggested in the figure, the perpendicular bisectors are asymptotic to each other in the downward direction.

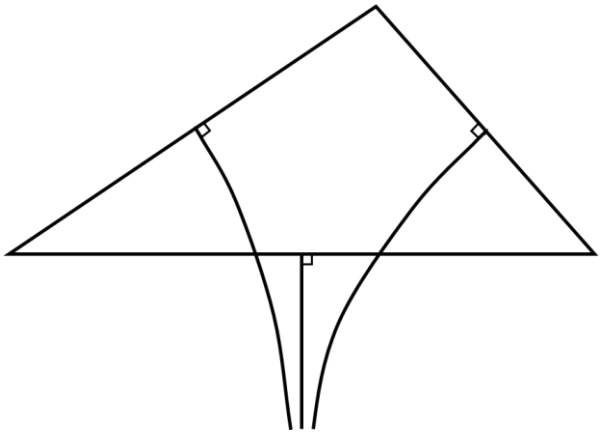


Figure 17. Perpendicular bisectors are boundary parallels

2.3.12 Gaussian Curvature

The curvature associated with various models of hyperbolic geometry varies. We mention "curvature" here because some of the formulas in hyperbolic geometry require the value of curvature for the specific model being used. The following is a high-level description of curvature [21]:

> **Gaussian curvature** is a powerful tool for characterizing surfaces, and it has a neat relationship with hyperbolic geometry:
>
> - Gaussian curvature captures how much a surface bends: It considers slices through a surface and their curvatures. A positive Gaussian curvature means the surface

curves in one direction (like a sphere), while a negative curvature indicates a saddle shape (like a hyperbolic paraboloid).

- Hyperbolic geometry is all about constant negative curvature: This geometry applies to surfaces where Gaussian curvature is always negative. In simpler terms, hyperbolic space is "curved outwards" everywhere.

Here's a breakdown of the connection:

- Negative Gaussian curvature = Hyperbolic point: If the product of a surface's principal curvatures (maximum and minimum curvatures at a point) is negative, that point has negative Gaussian curvature and is considered "hyperbolic."
- Surfaces with constant negative Gaussian curvature: These surfaces, known as pseudospheres, are prime examples of hyperbolic geometry. Every point on such a surface has constant negative Gaussian curvature.

In essence, Gaussian curvature helps identify regions in space that behave according to the rules of hyperbolic geometry. They both deal with how curved space is, but Gaussian curvature is a more general concept that can apply to any surface, whereas hyperbolic geometry focuses on spaces with a specific kind of curvature (constant and negative).

For example, the formulas for the circumference and area of a circle in hyperbolic geometry depend on the value of the Gaussian curvature associated with the particular model being used [22]. If r is the radius of a circle, and $R = \frac{1}{\sqrt{-K}}$ where K is the Gaussian curvature, then the circumference of a circle is given by the formula

$$2\pi R \sinh\left(\frac{r}{R}\right)$$

and the area of the circle is given by

$$4\pi R^2 \left(\sinh\left(\frac{r}{2R}\right)\right)^2$$

Note that $\sinh x$ is the hyperbolic sine [23] and is defined in terms of the exponential function, i.e.,

$$\sinh x = \frac{e^x - e^{-x}}{2}$$

Although hyperbolic geometry applies to any surface with a constant negative Gaussian curvature, it is usual to assume a scale in which the curvature $K = -1$.

2.3.13 Models of the Hyperbolic Plane

Thus far, we have not used a specific model of the hyperbolic plane, but rather, we have worked directly from the various postulates and theorems, with the use of some informal drawings. However, there are several models of the hyperbolic plane that completely satisfy the postulates of hyperbolic geometry.

One such model is the Klein model (which is a special case of something called the **Beltrami–Klein model** [24] in 2 dimensions). In the Klein model the interior of a disk (i.e., unit circle) is used to represent the entire hyperbolic plane. Lines are represented by chords of the unit circle. The points

on the boundary of the unit circle are called ideal points; although well defined, they do not belong to the hyperbolic plane.

Figure 18 shows three lines through point P that are parallel to line ℓ in the sense that they do no intersect ℓ.

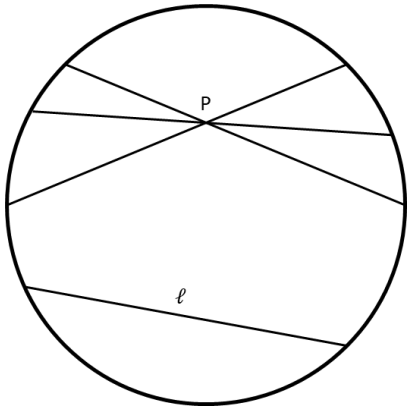


Figure 18. Lines through a point and parallel to a line – Klein model

Chords that meet on the boundary of the circle are boundary parallels to each other. All the lines (represented as chords) are boundary parallels to each other in Figure 19. The lines converge on point P which on the boundary of the circle and thus, outside the model (i.e., the lines do not intersect at P).

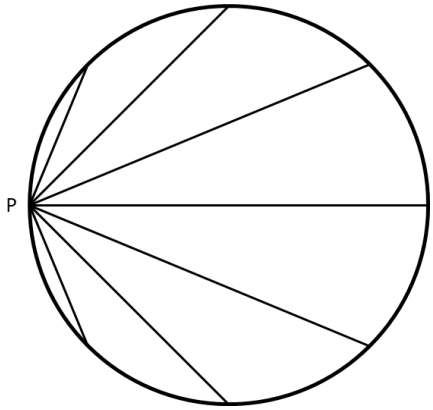


Figure 19. Boundary parallels

The **Poincaré disk model** [25], also known as the conformal disk model, is also based on the interior of the unit circle, but lines are represented by arcs of circles that are orthogonal (i.e., perpendicular) to the boundary of the circle, plus diameters of the boundary circle. Several lines are shown in Figure 20. For example, lines $\mathcal{h}, \ell$ and $\mathcal{n}$ are parallel to each other. Line $\mathcal{h}$ is a diameter. Line $\mathcal{m}$ intersects lines $\mathcal{h}$ and $\mathcal{n}$. Lines g and $\mathcal{h}$ are boundary parallels to each other. Since the boundary of the disk is not part of the model, g and $\mathcal{h}$ only approach (converge to) point Q. Each line in the drawing is perpendicular to the tangent line to the circle at two points, e.g., line ℓ is perpendicular to the tangent lines to the circle at points P and R.

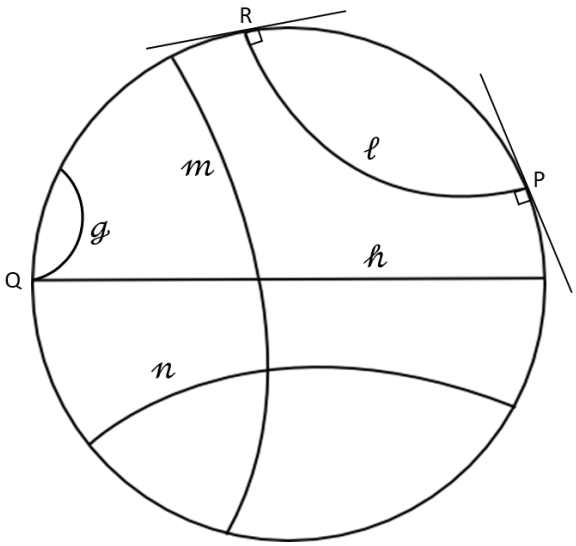


Figure 20. Example lines in the Poincaré disk model

Figure 21 depicts a collection of lines (represent as arcs) through the point P and parallel to line ℓ. (Figure 21 is a modification of a figure from the Wikipedia article "Poincaré disk model" [25].)

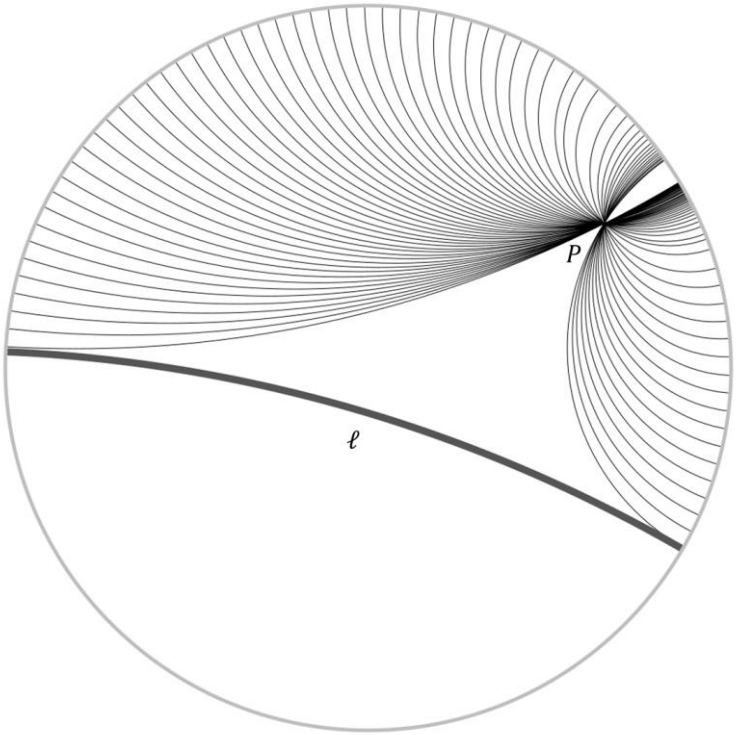


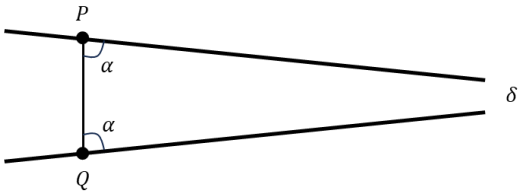
Figure 21. Lines through a point and parallel to a line – Poincaré disk model

There are several other models of hyperbolic geometry, see "Hyperbolic geometry: Models of the hyperbolic plane" [26].

2.3.14 Horocycles

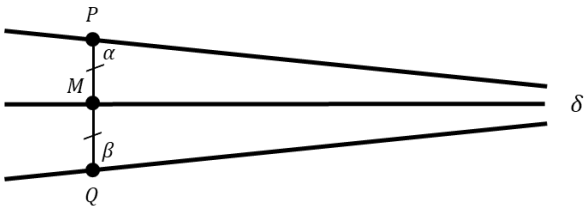
A **horocycle** is a continuous curve defined by a set of asymptotic (boundary) parallel lines, with one point of the horocycle coming from each line. The points of the horocycle are defined in a particular manner which we discuss below. [**Author's Remark**: Be patient, as it will take some development before we formally define a horocycle.]

Consider two asymptotic parallels $P\delta$ and $Q\delta$ as shown in the figure below. As indicated by the notation, P is on one line, Q is on the other line, and both lines have direction δ. If P and Q are such that the trilateral $PQ\delta$ is equiangular, then we say that P and Q are **corresponding points**. In the figure below, P and Q are corresponding points on asymptotic parallels $P\delta$ and $Q\delta$.



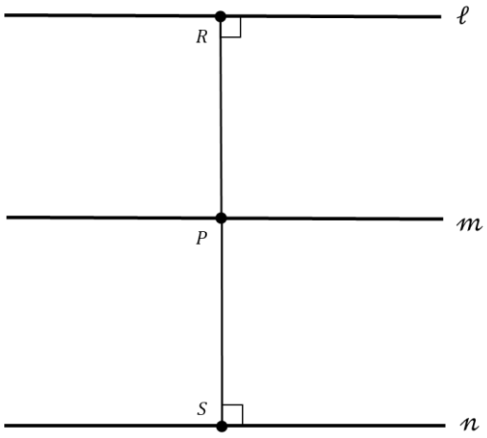
Theorem 49. Given two asymptotic parallel lines $P\delta$ and $Q\delta$. Points P and Q are corresponding points on these lines if and only if the perpendicular bisector of PQ is parallel to $P\delta$ and $Q\delta$ in the direction δ.

Proof: Assume that the perpendicular bisector of PQ is parallel to $P\delta$ and $Q\delta$ as shown in the figure below. By Theorem 40, trilateral $MP\delta$ and $MQ\delta$ are congruent, and so, $\alpha = \beta$ and in turn, P and Q are corresponding points.



Going in the other direction, assume P and Q are corresponding points. By Theorem 40, trilaterals $MP\delta$ and $MQ\delta$ are congruent. So, by symmetry, if line $P\delta$ meets the parallel bisector of PQ ($M\delta$ in the above figure) at a point (call it A), then so does line $Q\delta$, but this is a contradiction since we were given that $P\delta$ and $Q\delta$ asymptotic parallels to each other. Thus, neither $P\delta$ nor $Q\delta$ meet the parallel bisector of PQ, i.e., the perpendicular bisector of PQ is parallel to $P\delta$ and $Q\delta$ in the direction δ. ■

In the context of the following theorems, a line m is said to be equidistant from lines ℓ and n, if the perpendicular (shortest) distance from any point P on m to either lines ℓ or n is the equal. In the figure below, $PR = PS$. To be clear, the distance from line m to the other two lines can vary depending on the point chosen on m (keep this in mind when reading the following theorem).

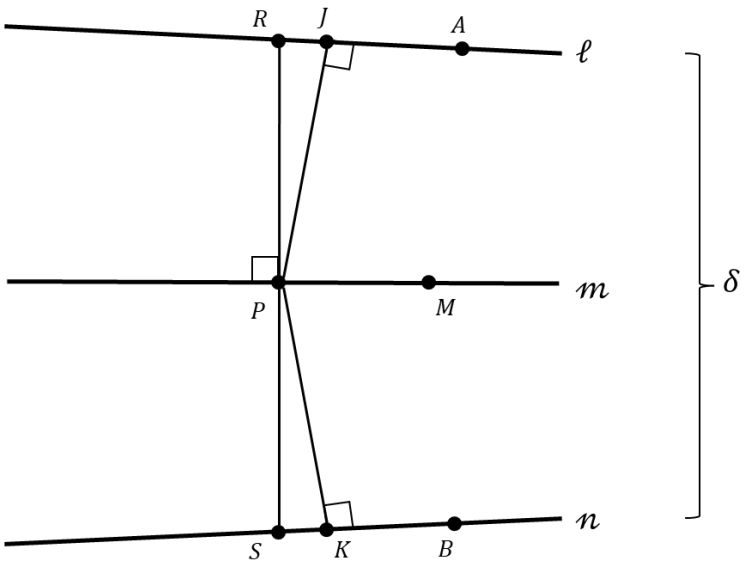


Theorem 50. Given two asymptotic (i.e., boundary) parallels $P\delta$ and $Q\delta$, there exists a line $M\delta$, each of whose points is equidistant from and parallel to the two given lines.

Proof: See Theorem 2 in Chapter V, Section 2 of Gans [11].

Theorem 51. Given two asymptotic (i.e., boundary) parallels $A\delta$ and $B\delta$, and a point R on $A\delta$, there exists a unique corresponding point on $B\delta$.

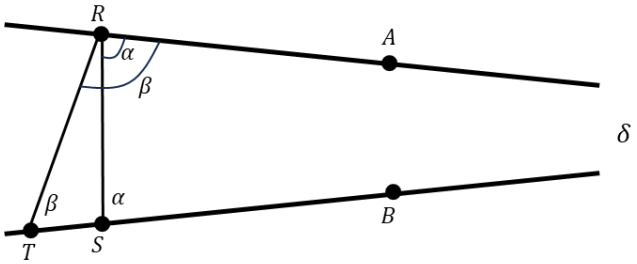
Proof: By Theorem 50, there exists a line m which is equidistant between lines $A\delta$ and $B\delta$ (labeled as lines ℓ and n in the figure below). Let P be the projection of R onto line m, and let J on ℓ and K on n be equidistant from P. Select the point S on line n such that $SK = RJ$. By the SAS triangle congruence principle, triangle RJP and SKP are congruent, which implies that $\angle PRJ = \angle PSK$ and $\angle RPJ = \angle SPK$.



By Theorem 40, trilaterals $AJPM$ and $BKPM$ are congruent, and so, $\angle JPM = \angle KPM$. Since $90° = \angle RPM = \angle RPJ + \angle JPM$, $\angle JPM = \angle KPM$ and $\angle RPJ = \angle SPK$, we have that $\angle SPM = 90°$. Thus, R, P and S are collinear. From the collinearity of R, P and S and $\angle PRJ = \angle PSK$, we have that $RS\delta$ is an equiangular trilateral, and thus, R and S are corresponding points.

. . .

To prove uniqueness, assume there are two points on n corresponding to R, i.e., points S and T. Further, assume the interior angles of equiangular trilateral $RS\delta$ are of measure α, the interior angles of equiangular trilateral $RT\delta$ is β and $\beta > \alpha$ (with loss of generality). See the figure below.



Since the exterior angle of a triangle is greater than either of the non-adjacent interior angles (by Book I, Prop. XVI of Euclid's Elements), we have that $\alpha = \angle RSB > \angle RTS = \beta$ which contradicts our earlier assumption. Thus, our assumption of two corresponding points to R is false. ■

Just a few more theorems before we get to the definition of a horocycle …

Theorem 52. If points P, Q and R lie on three asymptotic parallels $P\delta$, $Q\delta$ and $R\delta$ such that P and Q are corresponding points, and Q and R are corresponding points, then P, Q and R are noncollinear.

Proof: See Theorem 2 in Section 57 of Wolfe [10].

Theorem 53. If points P, Q and R lie on three asymptotic parallels $P\delta$, $Q\delta$ and $R\delta$ such that P and Q are corresponding points, and Q and R are corresponding points, then P and R are corresponding points.

Proof: See Theorem 3 in Section 57 of Wolfe [10].

Consider a line ℓ in direction δ and point P on it. For each asymptotic parallel n, there exists a unique point Q on n that corresponds to P (see Figure 22). The locus consisting of P and all such corresponding points (for each asymptotic parallel to ℓ) is a **horocycle**. The horocycle is determined by line ℓ, point P and direction δ. The asymptotic parallels to ℓ in the direction δ, and ℓ itself, are known as the **radii of the horocycle**. Since the given horocycle is determine by $P\delta$, we denote it with the notation (P, δ).

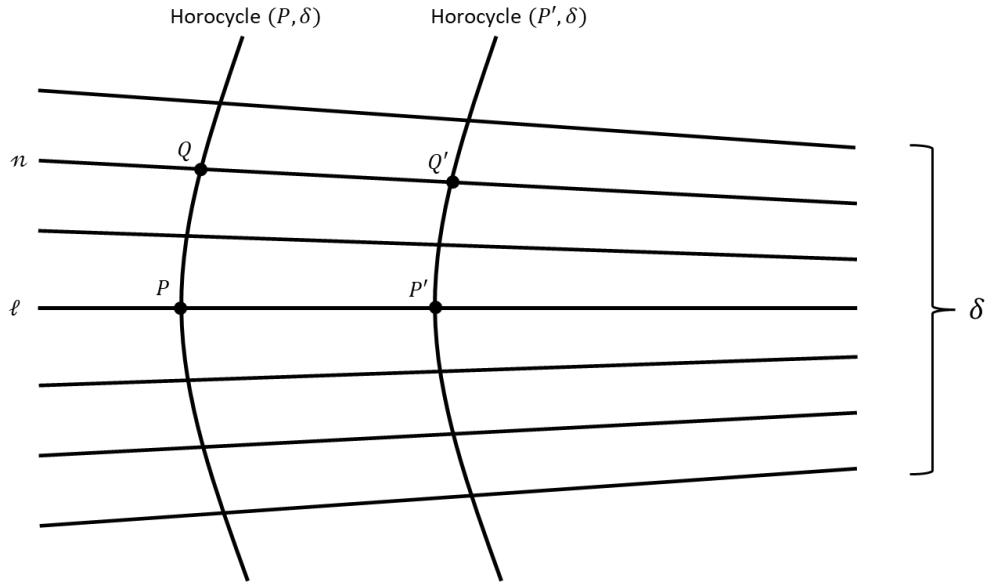


Figure 22. Horocycles

It follows by Theorem 53 that all the points on a horocycle are mutually corresponding points. Thus, a horocycle can be determined by any of its points. In Figure 22, (P, δ) and (Q, δ) refer to the same horocycle. If P' is another point on ℓ, then horocycle (P', δ) is a different from horocycle (P, δ), but they both have the same direction and the same radii. Distinct horocycles which have the same direction, and hence the same radii, are referred to as **codirectional horocycles**. For example, (Q, δ) and (Q', δ) are codirectional horocycles.

Horocycles are not circles but they share several characteristics with circles, e.g.,

- There exists a unique circle, with a given center, which passes through a given point. In comparison, there exists a unique horocycle with a given direction, which passes through a given point.
- Two concentric circles share no points. In comparison, codirectional horocycles share no points.
- A unique radius is associated with each point on a circle. A unique radius is associated with each point on a horocycle.
- A tangent to a horocycle at a point on the horocycle is defined to be the line through the point which is perpendicular to the radius associated with the point. This is the same as the definition of a tangent to a circle.

Horocycles appear differently depending on the model being used for the hyperbolic plane. Figure 23 (adapted from the Wikipedia article Horocycle [27]) depicts a horocycle using the Poincaré disk model.

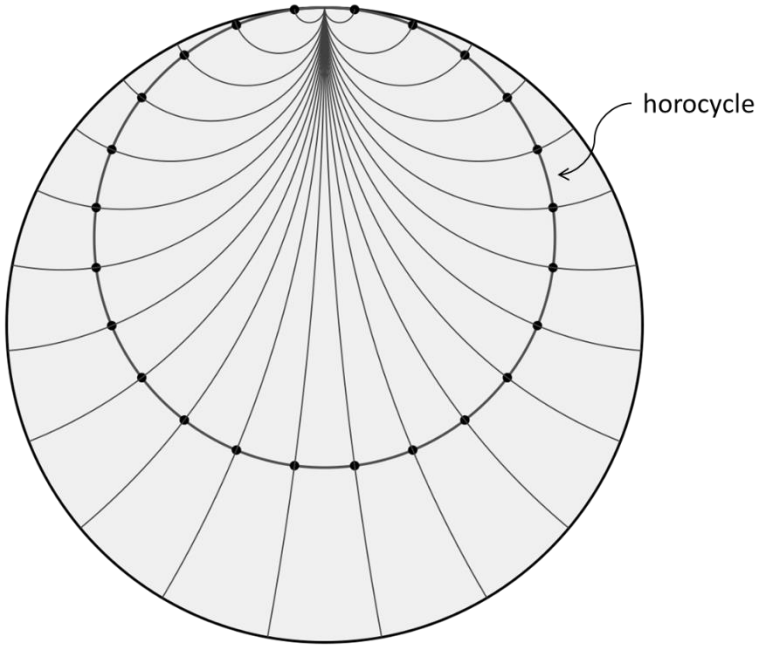


Figure 23. Horocycle in Poincaré disk model

2.4 Elliptic Geometry

2.4.1 Overview

As noted previously, elliptic geometry is an example of a geometry in which Euclid's parallel postulate does not hold. There are two variations of elliptic geometry i.e., double elliptic geometry and single elliptic geometry, both of which have no parallel lines. In double elliptic geometry, each pair of lines intersect in exactly two points. In single elliptic geometry, each pair of lines intersect in exactly one point. In both types of elliptic geometry, straight lines are finite in length.

Recall that the basis for hyperbolic geometry was to replace Euclid's parallel axiom by the contrary statement or more precisely, a logical equivalent of the parallel postulate. From there, we were able to make use of Propositions 1 through 28 from Book I of Euclid's Elements. Unfortunately, no such simple approach is possible for elliptic geometry, since the axioms used in proving many of the first 28 propositions in Book I of Euclid's Element involve assumptions that do not hold in elliptic geometry. In the interest of simplifying the introduction to this topic, the axiomatic foundations for elliptic geometry are not discussed in this book. For the interested reader, Chapter VII, Section 7 of Gans [11] provides an axiomatic presentation for double elliptic geometry, and Chapter VIII, Section 4 of Gans [11] provides an axiomatic presentation for single elliptic geometry. There is also a more technical discussion of the axioms for elliptic geometry in the article "Axioms for Elliptic Geometry" [28].

In lieu of an axiomatic development, we discuss specific models that meet the axioms of elliptic geometry. In the case of double elliptic geometry, a spherical model is used. In the case of single elliptic geometry, a hemispherical model is used.

The qualifier "elliptic" is a misnomer to some extent. From a footnote on Page 94 of Coxeter [29]:

> The name "elliptic" is possibly misleading. It does not imply any direct connection with the curve called an ellipse, but only a rather far-fetched analogy. A central conic is called an ellipse or a hyperbola according as it has no asymptote or two asymptotes. Analogously, a non-Euclidean plane is said to be elliptic or hyperbolic according as each of its lines contains no point at infinity or two points at infinity.

In what follows, we provide a description of the two variants of elliptic geometry. Proofs of theorems are not provided.

2.4.2 Double Elliptic Geometry

In the spherical model of double elliptic geometry, the surface of a sphere represents all the points in the geometry.

A **great circle** is the circular intersection of a sphere and a plane passing through the sphere's center point. Great circles are the straight lines of spherical geometry, and as such, straight lines in spherical geometry are closed curves of finite length. If the radius of the sphere (and thus, also the great circle is r), then the length of each great circle on the sphere is $2\pi r$.

Two great circles meet in two points known as **antipodal points**, i.e., points at the ends of a diameter of the sphere. In Figure 24, $\mathcal{A}$ and $\mathcal{B}$ are great circles who intersect at antipodal points P and P'. The center of the sphere as well as the center of the two great circles is point O. All great circles that go through a given point will meet again at the antipodal point of the given point.

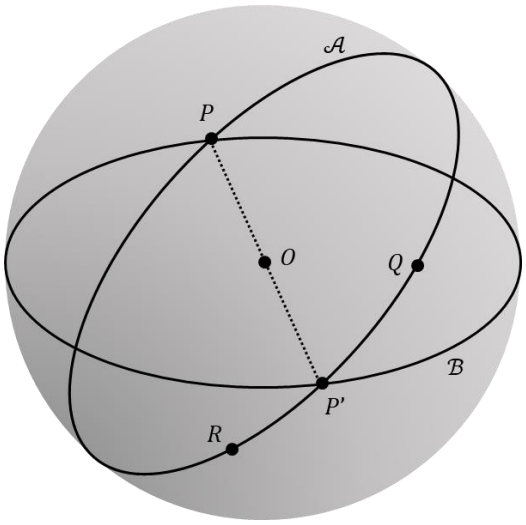


Figure 24. Sphere and great circles

We have the following facts about points and straight lines (i.e., great circles) in spherical geometry:

- Each pair of straight lines meet in two points known as antipodal points.
- Through each pair of antipodal points passes an infinite number of straight lines. Through each pair of non-antipodal points passes a unique straight line. For example, $\mathcal{A}$ is the only great circle passing through non-antipodal points Q and R in Figure 24.
- Through each point there pass infinitely many straight lines, the totality of which covers the entire sphere.

. . .

The shortest path on a sphere joining one point to another is known as a **geodesic arc** on the sphere.

- For non-antipodal points, there is also a second (longer path) between the points, but this path is not considered to be a geodesic arc. In the case of antipodal points, there are two shortest paths between the points.
- A geodesic arc is always an arc of a unique great circle.
- A geodesic arc is the analog of a line segment in Euclidean geometry.

The distance between two non-antipodal points on a sphere is defined to be the length of the geodesic arc joining them (as computed using Euclidean geometry applied to a sphere). If the two points are antipodal, then the distance between them is half the circumference of a great circle, i.e., πr.

Theorem 54. Three points on a sphere necessarily lie on the same great circle, if two of them are antipodal.

Proof: There are an infinite number of great circles with the same two antipodal points (think of the North and South poles on a globe and the meridian lines). It is just a matter of selecting the great circle that contains the third point. ■

Consider three distinct points not on the same great circle. By Theorem 54, each two of the points must be non-antipodal and lie on a unique geodesic arc.

. . .

If A, B, C are any three points on a sphere, then $AB + BC > AC$, where the notation AB denotes the distance between A and B. [We also use the notation AB to denote the geodesic arc between A and B.] This relationship between three points is known as the triangle inequality [30].

The sum of the angles of a spherical triangle is greater than $180°$ and less than $540°$. Two spherical triangles are said to be congruent if their corresponding sides and angles are equal. Two spherical triangles are not necessarily congruent if two angles and the side opposite one are equal, i.e., the AAS triangle congruence principle does not necessarily hold in spherical geometry (see Case 5 in the "Oblique triangles" section of the Wikipedia article "Spherical trigonometry" [31]). The other triangle congruence principles do hold true, i.e., SSS, SAS, ASA and AAA.

The area of a spherical triangle is given by the formula

$$r^2(\alpha + \beta + \gamma - \pi)$$

where r is the radius of the sphere, and α, β, γ are the measures of the angles in the triangle (in radians).

For a right triangle on a sphere of radius r (as shown in Figure 26), we have the formulas listed below. These are known as Napier's rules for right spherical triangles [31].

$$\sin\alpha = \frac{\sin\left(\frac{a}{r}\right)}{\sin\left(\frac{c}{r}\right)}, \qquad \cos\alpha = \frac{\tan\left(\frac{b}{r}\right)}{\tan\left(\frac{c}{r}\right)}, \qquad \tan(\alpha) = \frac{\tan\left(\frac{a}{r}\right)}{\sin\left(\frac{b}{r}\right)}$$

$$\cos\left(\frac{a}{r}\right)\cos\left(\frac{b}{r}\right) = \cos\left(\frac{c}{r}\right)$$

Similar formulas hold for the angle β. Further, it is possible for a triangle to have 0,1,2 or 3 right angles in spherical geometry.

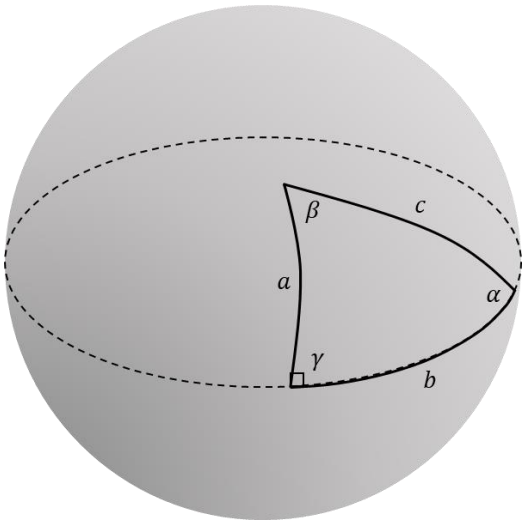


Figure 25. Right triangle in spherical geometry

We also have the **spherical law of sines** [32], i.e.,

$$\frac{\sin\left(\frac{a}{r}\right)}{\sin\alpha} = \frac{\sin\left(\frac{b}{r}\right)}{\sin\beta} = \frac{\sin\left(\frac{c}{r}\right)}{\sin\gamma}$$

and the **spherical law of cosines** [33], i.e.,

$$\cos\left(\frac{a}{r}\right) = \cos\left(\frac{b}{r}\right)\cos\left(\frac{c}{r}\right) + \sin\left(\frac{b}{r}\right)\sin\left(\frac{c}{r}\right)\cos\alpha$$

The Wikipedia article “Spherical trigonometry” [31] has an extensive list of formulas for spherical trigonometry.

. . .

All the great circles which are perpendicular to a given great circle $\mathcal{A}$ meet in two antipodal points known as the **poles** of $\mathcal{A}$. In Figure 26, the great circles perpendicular to great circle $\mathcal{A}$ all meet at poles N and S. In the figure, the great circles perpendicular to $\mathcal{A}$ are the vertical curved lines, e.g., $\mathcal{B}$ and $\mathcal{C}$. In terms of visualization, it may help to think of $\mathcal{A}$ as the equator on a globe, the great circles perpendicular to $\mathcal{A}$ as being longitudinal (aka meridian) lines, and N and S as being the North and South poles. [**Author’s Remark**: the horizontal curved lines are not great circles (except for $\mathcal{A}$). They appear here since I reused some clipart for my drawing.]

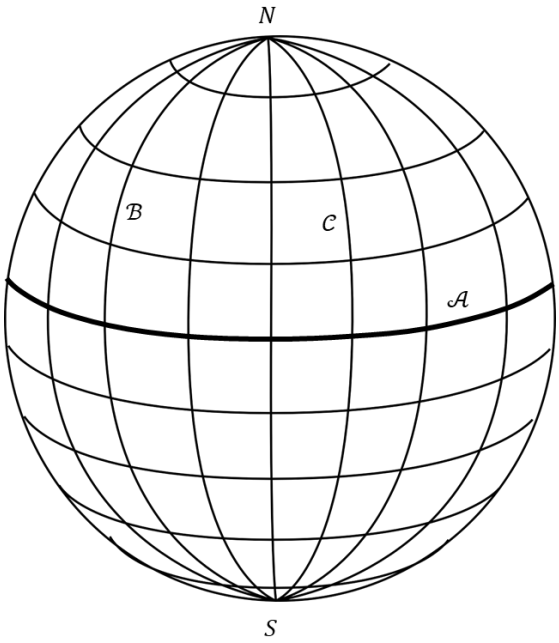


Figure 26. Great circles perpendicular to a given great circle meet at two points

2.4.3 Single Elliptic Geometry

For single elliptic geometry, we use what is called a **modified hemisphere model**. Each pair of antipodal points on the boundary of the hemisphere are considered to be the same point (see Figure 27). In this model, semicircles are effectively converted into closed curves known as **modified curves**. The modified curves represent straight lines. For example, AEA, CEC, ACA and CFC are modified curves in the figure.

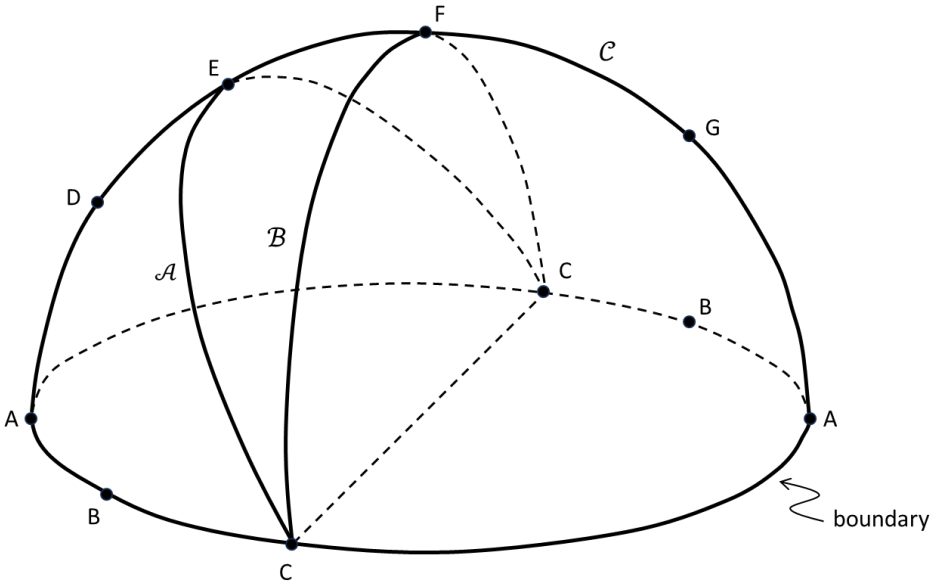


Figure 27. Modified hemisphere model

In single elliptic geometry, two straight lines intersect in just one point. If we did not make the modification to the hemisphere (equating antipodal points), some pairs of straight lines would have

two points of intersection and others would have only one point of intersection. For example, semicircles $\mathcal{A}$ and $\mathcal{B}$ in Figure 27 would have two points of intersection if we did not equate antipodal points, whereas semicircles $\mathcal{A}$ and $\mathcal{C}$ have one point of intersection (whether or not we equate antipodal points).

Assume the radius of the hemisphere is of measure r.

The boundary of the hemisphere is a circle. Its circumference is πr rather than $2\pi r$, since the boundary is completely traversed by going continuously from any point on the boundary to its antipodal point (which is a distance πr). The boundary is also considered a modified curve. The other modified curves (being semicircles) also have circumference (or more precisely, semi-circumference) πr.

Some additional properties of modified curves are as follows:

- The shortest path (on the modified hemisphere) joining two points of the modified hemisphere must be an arc of a modified curve, known as a **geodesic arc** (same term that we used for spherical geometry).
- Through each point (on the modified hemisphere) there pass infinitely many modified curves, the totality of which cover the entire modified hemisphere.
- Through each pair of points (on the modified hemisphere) there passes a unique modified curve.
- A pair of modified curves always meet at a unique point.
- A pair of points are said to be opposite if they divide their associated modified curve (i.e., straight line) into equal parts. In Figure 27, A and F are opposite points, and A and D are non-opposite points. Conversely, if a pair of points divide their associated modified curve into two equal parts, then the points are opposite.
- Opposite points are joined by exactly two geodesic arcs of the same length $\frac{\pi r}{2}$, and non-opposite points are joined by exactly one. Keep in mind that "geodesic arc" implies shortest distance.
- The maximum distance between two points on the modified hemisphere is $\frac{\pi r}{2}$.
- All modified curves perpendicular to a given modified curve meet in a unique point.

. . .

Triangles are a bit strange in hemispherical geometry. For example, consider the situation shown in Figure 28. Let T_1 be the triangle with sides ADF, AC (labelled as x) and CEF, and T_2 be the triangle with sides ADF, AC (labelled as y), and AF (on the right-side of the hemisphere). While it looks disconnected, T_2 is a triangle since the two appearances of point C coincide based on our definition of the modified hemisphere. Even more strange is that triangles T_1 and T_2 have corresponding sides of equal length but are not congruent since the angle at F differs in the two triangles (with one angle being the supplement of the other).

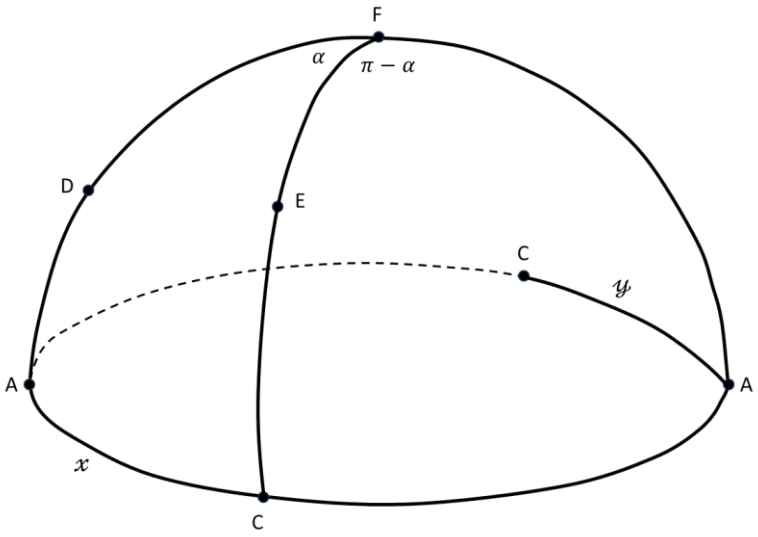


Figure 28. Triangle example in hemispherical geometry

The triangle inequality that we mentioned for spherical geometry also holds true for hemispherical geometry.

3 Topology

Point set topology is a disease from which the human race will soon recover. – Henri Poincare

[Author's Remark: So wrong … even the great ones make big mistakes.]

3.1 Overview

In this section, we provide a summary of the mathematical topic known as **topology**. Topology is the study of shapes and spaces. What happens if one allows geometric objects to be stretched or squeezed but not broken? In fact, there's quite a bit of structure in what remains, which is the principal topic of study in topology. From the perspective of topology, a coffee cup and an donut (or torus, to use the mathematical term) are equivalent since one can be continuous deformed into the other without breaking, see the demonstration in the YouTube video "Intro to Topology - Turning a Mug Into a Doughnut" [34].

Some additional definitions of topology:

- From the Wikipedia article on topology [35]: In mathematics, topology (from the Greek words τόπος, 'place, location', and λόγος, 'study') is concerned with the properties of a geometric object that are preserved under continuous deformations, such as stretching, twisting, crumpling, and bending; that is, without closing holes, opening holes, tearing, gluing, or passing through itself.
- From The Free Dictionary by Farlex: Topology is the study of certain properties that do not change as geometric figures or spaces undergo continuous deformation. These properties include openness, nearness, connectedness, and continuity.
- From the Wolfram MathWorld article on topology [36]: Topology is the mathematical study of the properties that are preserved through deformations, twistings, and stretchings of objects. Tearing, however, is not allowed. A circle is topologically equivalent to an ellipse (into which it can be deformed by stretching) and a sphere is equivalent to an ellipsoid.

A **topological invariant** is something immutable about a shape, no matter how we stretch and deform it. More formally, a topological invariant is a property of a topological space (a mathematical object that captures the basic idea of "shape") that is preserved under a homeomorphism. A **homeomorphism** is a continuous function between two spaces that is also invertible (meaning there's a corresponding function going the other way). Intuitively, a homeomorphism is a way of continuously deforming one space into another. Some examples of topological invariants are as follows:

- Connectedness: Whether a space is in one piece or multiple pieces.
- Compactness: Whether every cover of a space (a collection of regions that include all the space and possibly more) has a finite subcover. This concept is a bit different to grasp without further development (which we will provide in Sections 3.4.6 and 3.5.7). Compactness is a property that seeks to generalize the notion of a closed and bounded subset of Euclidean space.
- Genus: The number of "holes" in a surface (like the hole in a donut or torus).

We will address these invariants and others in the following subsections.

In the next subsection, we provide a relatively simple example involving the classification of the letters of the alphabet based on the number of holes, and the number and types of vertices. In Section 3.3, we consider surfaces (a 2-dimensional, connected shape such as the boundary of a sphere). In Section 3.4, a generalized type of geometry known as a metric space is studied. A metric space is a set with a distance defined between each pair of elements in the set. Many of the properties of metric spaces can be defined without the concept of distance, which leads us to topological space (discussed in Section 3.5). We conclude the discussion of topological spaces with a more formal description of surfaces (see Section 3.5.9).

3.2 A First Example: Classifying the Letters of the Alphabet

From a topological viewpoint, what letters of the alphabet are equivalent (homeomorphic to use the topological term)? To answer this question, we first need to select a specific alphabet. For the task at hand, we choose the alphabet used for the English language, and the Calibri font (as shown below). Calibri was selected since its letters have no serifs which makes the classification task a bit easier.

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Two letters are considered homeomorphic if one letter can be bent, stretched or compressed to form the other letter. For example, we can bend the letter C to form a Z. So, C and Z are homeomorphic in the topology for this problem.

The are two types of invariants in our topology, i.e.,

- Homeomorphic letters must have the same number of holes. For example, D and O each have the same number of holes. Their difference in the shape of their holes does not matter in our topology.
- The other topological invariant is the number and type of vertices in a letter. In this context, a vertex is a point where multiple lines intersect. The number of intersecting lines at a point determines the vertex type of the point. For example, the letter T has a vertex where three lines intersect (this is known as a 3-vertex). The letters K and X each have a 4-vertex. It is not allowed (in this topology) to change the vertex type of a point, e.g., we cannot morph a T into an L. Each letter has an infinite number of 2-vertices, and some letters have 1-vertices at their tips, e.g., the letter “I” has a 1-vertex at its top and bottom.

Another method to determine the type of vertex is to enclose the vertex in question with a circle (neighborhood) that includes the vertex in question but no other vertices of type 3 and above. Next, remove the vertex and then count the number of separate lines. This technique is applied to the letters R and Q, as shown in Figure 29. Drawing the circle to isolate the point is critical. For example, if we just removed the vertex in question from Q, we would be left with only two components which would give us the wrong answer for the associated vertex type.

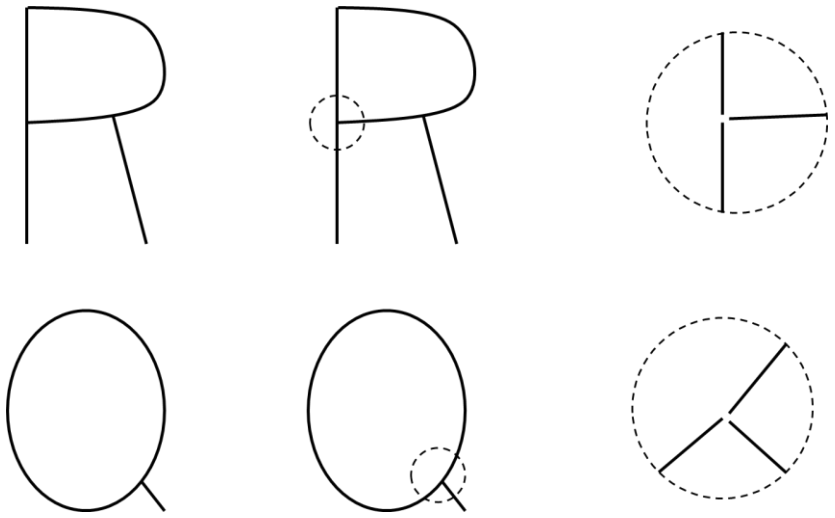


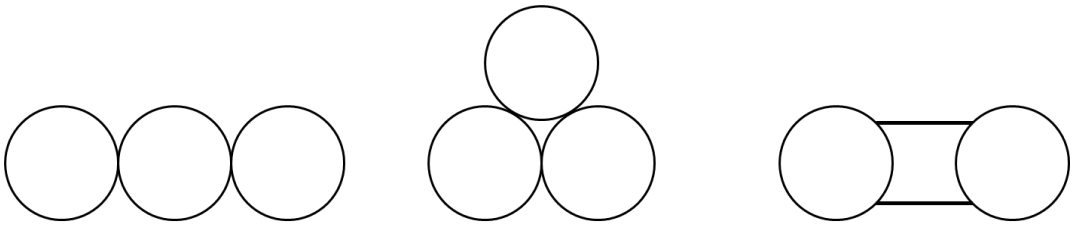
Figure 29. Determination of vertex type for R and Q

Using the above methodology for distinguishing letters, we get the classification of homeomorphic letters shown in Table 1. This methodology is based on a technical paper by Vasanthi [37].

Table 1. Topological classification of letters

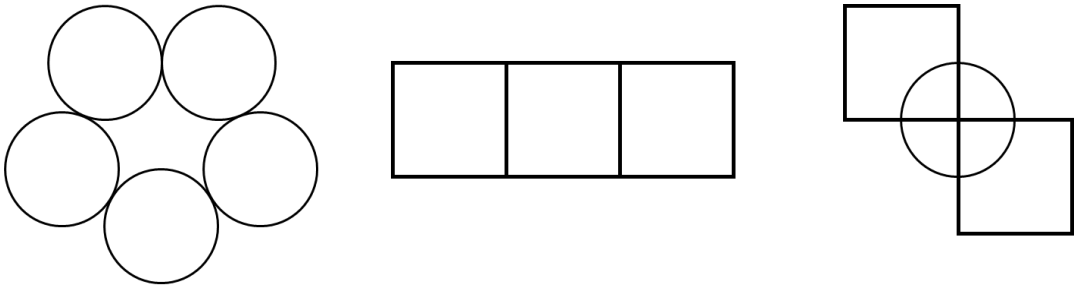
Invariants	**Homeomorphic Letters**
0 holes, all points are 2-vertices (except for the tips of the letters, which are 1-vertices)	C G I J L M N S U V W Z
1 hole, all points are 2-vertices	D O
0 holes, one 3-vertex	E F T Y
1 hole, one 3-vertex	P Q
0 holes, one 4-vertex	K X
0 holes, two 3-vertices	H
1 hole, two 3-vertices	A R
2 holes, two 3-vertices	B

This methodology can be extended to more general configurations beyond the letters of the alphabet, see the examples in the figure below. The middle configuration is a bit tricky, i.e., the gap between the three circles is also a hole.



3 holes, two 4-vertices 4 holes, three 4-vertices 3 holes, four 3-vertices

The reader is challenged to classify the following configurations based on the number of holes and vertex types.



Our analysis of the alphabet and the generalized configurations assume the figures are 1-dimensional configurations embedded in 2-dimensions. If we allow for the 1-dimensional closed curves to be embedded in 3-dimensions, we enter the world of mathematical knots. Knots have been studied extensively, see Section 8 of Mathematical Vignettes II [38].

3.3 Surfaces

3.3.1 Introduction

In topology, a surface [39] is a two-dimensional manifold. By “two-dimensional”, we mean that a surface has no thickness. The qualifier “manifold” means that that there exists a neighborhood (small area) around each point of the surface that is homeomorphic to 2-dimensional Euclidean space $\mathbb{R}^2$.

Some surfaces arise as the boundaries of three-dimensional solid figures, e.g., the surface of a sphere or the surface of a polyhedron (a three-dimensional shape with flat polygonal faces, straight edges and sharp corners or vertices). The five platonic solids [40] are examples of polyhedra, see Figure 30. The platonic solids are the only regular polyhedra, where “regular” means the faces are congruent (identical in shape and size) regular polygons (all angles congruent and all edges congruent), and the same number of faces meet at each vertex.

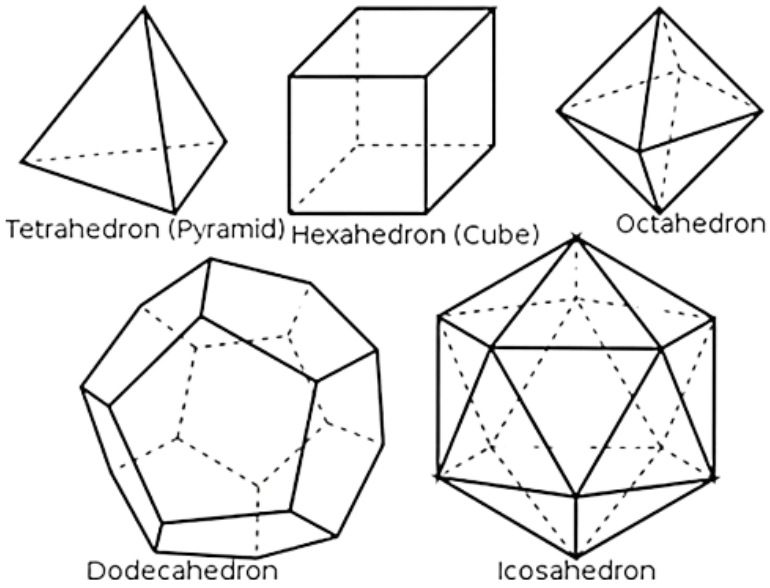


Figure 30. Platonic solids

In other cases, surfaces arise as graphs of functions of two variables. Figure 31 depicts the surface generated by the equation

$$z = \frac{x^2}{4} + \frac{y^2}{7}$$

This surface extends indefinitely in the direction of the positive z-axis.

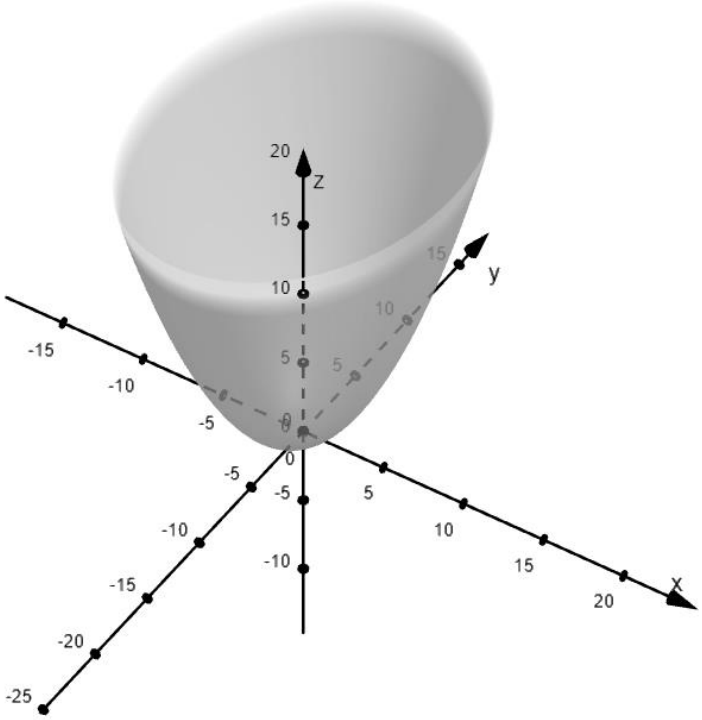


Figure 31. Surface defined by an equation

Our goal in this section is to classify surfaces, i.e., determine criteria that allows one to determine whether two given surfaces are homeomorphic (equivalent).

3.3.2 Euler's Characteristic

The various features of a polyhedron are labeled in Figure 32. The particular example is that of a pentagonal prism. A pentagonal prism has 7 faces, 10 vertices and 15 edges.

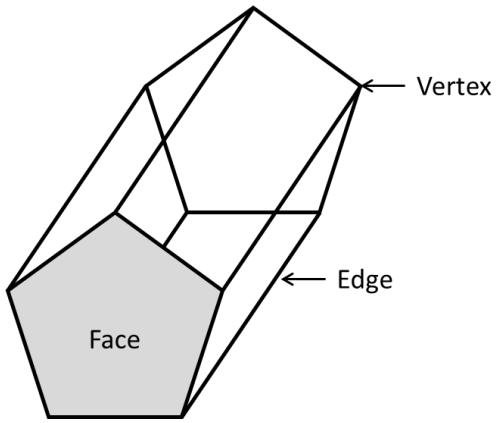


Figure 32. Features of a polyhedron

(The plural form of polyhedron is polyhedra or polyhedrons.) The history of polyhedra dates back millennia. From the Wikipedia article Polyhedron [41], we have the following historical background:

> Polyhedra appeared in early architectural forms such as cubes and cuboids, with the earliest four-sided Egyptian pyramids dating from the 27th century BC. The Moscow Mathematical Papyrus from approximately 1800–1650 BC includes an early written study of polyhedra and their volumes (specifically, the volume of a frustum). The mathematics of the Old Babylonian Empire, from roughly the same time period as the Moscow Papyrus, also included calculations of the volumes of cuboids (and of non-polyhedral cylinders), and calculations of the height of such a shape needed to attain a given volume.
>
> The Etruscans preceded the Greeks in their awareness of at least some of the regular polyhedra, as evidenced by the discovery of an Etruscan dodecahedron made of soapstone on Monte Loffa. Its faces were marked with different designs, suggesting to some scholars that it may have been used as a gaming die.

Given the long history of polyhedra, it is surprising (at least to the author of this book) that a very fundamental relationship between the number of faces, edges and vertices went undiscovered until the 16th century AD. Rather than just state the result, perhaps the reader can figure out the relationship among the faces, edges and vertices of a polyhedron, given a few hints. The first hint is given in Table 2. It is recommended that the reader verify the counts for a few of the items in the table.

Table 2. Number of vertices, edges and faces for several

Name	**Vertices *V***	**Edges *E***	**Faces *F***
Tetrahedron	4	6	4
Cube	8	12	6
Octahedron	6	12	8
Dodecahedron	20	30	12
Icosahedron	12	30	20
Pentagonal Prism In Figure 32	10	15	7

As a second hint, the relationship between the number of vertices, edges and faces is linear, i.e., has the form $aV + bE + cF = d$ where a, b, c, d are constants.

Last hint: $d = 2$.

The relationship is $V - E + F = 2$ and it is known as Euler's characteristic. According to the Wikipedia article "Euler characteristic":

> The Euler characteristic was originally defined for polyhedra and used to prove various theorems about them, including the classification of the Platonic solids. It was stated for Platonic solids in 1537 in an unpublished manuscript by Francesco Maurolico. Leonhard Euler, for whom the concept is named, introduced it for convex polyhedra more generally but failed to rigorously prove that it is an invariant.

The Euler characteristic for connected planar graphs follows the same formula as for polyhedral surfaces. There is one nuance, i.e., the exterior of the graph counts as one face (see the labeling of faces in the graph on the left of Figure 33). The formula holds even for connected planar graphs that do not have any closed loops, e.g., the graph on the right of Figure 33.

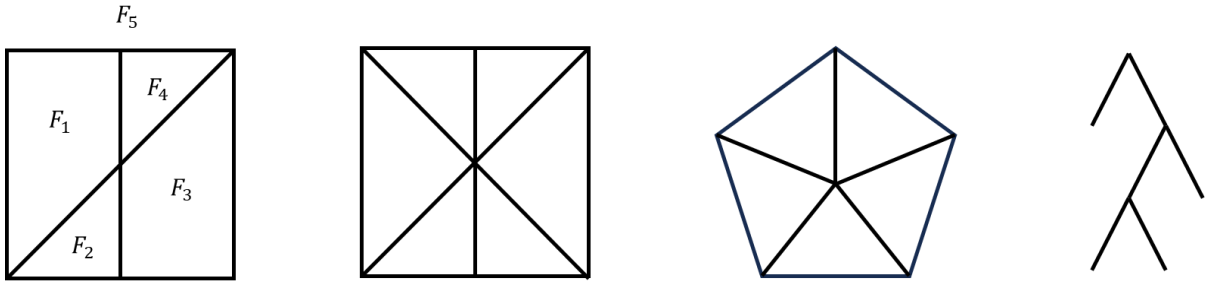


Figure 33. Euler's characteristic for connected planar graphs

Theorem 55. The number of vertices (V), edges (E) and faces (F) of a connected planar graph is given by the formula $V - E + F = 2$.

Proof: We will use proof by induction on the number of faces F in the planar graph G.

If $F = 1$, then G cannot have any cycles (otherwise the interior and exterior of the cycle would be 2 distinct faces). So, in this case, G no cycles and is by definition a tree. From a theorem in graph theory, we must have that the number of vertices in a tree is one more than the number of edges, i.e., $V = E + 1$ (see Proposition 4.2.4 in Levin [43]). Thus,

$$V - E + F = (E + 1) - E + 1 = 2$$

which proves the result for the case $F = 1$.

Assume the theorem is true for the case $F = k$ and let G be a connected planar graph with $k + 1 \geq 2$ faces. Since trees have only one face, G cannot be a tree and thus, must have a cycle (closed loop of edges). Choose any edge e in the cycle and remove it (but not its vertices) from G, forming a new graph $H = G - \{e\}$.

Clearly, the number of edges for graph H is $E(H) = E(G) - 1$ and the number of vertices is $V(H) = V(G)$.

Further, H has one less face than G, i.e., $F(H) = F(G) - 1 = k$ since the edge e being part of a cycle must separate two faces of G, which are joined into one face of H. By a basic result from graph theory (see the StackExchange Mathematics article [44]), H is connected, and so, our inductive hypothesis applies to H. Thus,

$$2 = V(H) - E(H) + F(H) = V(G) - (E(G) - 1) + (F(G) - 1) = V(G) - E(G) + F(G)$$

which completes the inductive step. ■

In terms of polyhedra, we have the following theorem concerning Euler's characteristic.

Theorem 56. For any convex polyhedron, and more generally to any polyhedron whose boundary is topologically equivalent to a sphere and whose faces are topologically equivalent to disks, the relationship between the number of vertices, edges and faces is given by the formula $V - E + F = 2$.

Proof: We only provide a sketch of a proof here. For several detailed proofs, see Eppstein [45].

The general idea is quite simple, i.e., take a given polyhedron, remove one of its faces and then flatten the remaining figure into a planar graph. The assumption that the polyhedral surface is homeomorphic (equivalent) to the sphere is what makes this possible. The flattened figure has one less face than the original but recall that for planar graphs we count the exterior region as a face, which brings us back to the same number of faces as the original polyhedron. Thus, we can apply Theorem 55 to get the desired result. ■

At this point, the reader may be thinking "to what extent does this formula apply?"

For example, consider the hexagon torus in Figure 34. It has 24 vertices, 48 edges and 24 faces, and thus, $V - E + F = 0$. In fact, all toroidal polyhedra (i.e., polyhedra with one hole) have the value of their Euler's characteristic equal to 0. The Wikipedia article "Toroidal polyhedron" [46] provides many examples.

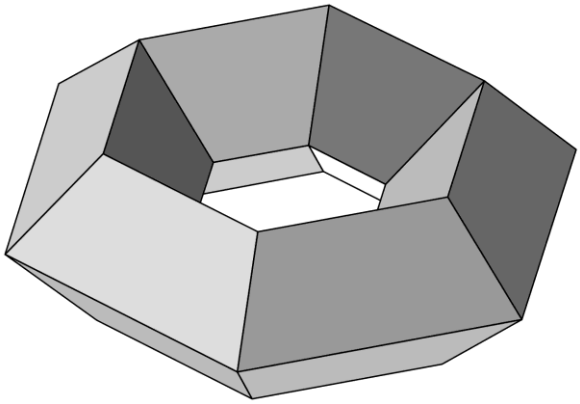


Figure 34. Hexagonal torus

The hexagonal torus approximates a torus. We can extend the idea by using more faces and come even closer to the shape of a torus, see Figure 35. As we extend the approximation, the Euler characteristic continues to be 0. Taking the limit, we arrive at a torus, and conclude (at least, intuitively) that the Euler characteristic of a torus is 0.

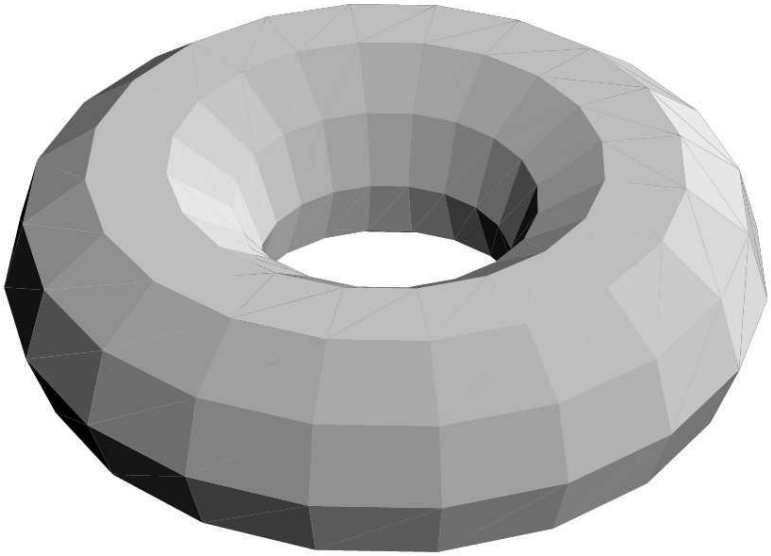


Figure 35. Approximating a torus with a toroidal polyhedra

Another way of computing the Euler characteristic of a torus is illustrated in Figure 36.

- We start with a rectangle having four vertices, four edges and two faces (interior and exterior of the rectangle).
- We bend the rectangle into the cylinder, as shown in the center of the figure. Two of the edges are joined to become one, and two pairs of vertices are joined together.
- Finally, we stretch and bend the cylinder into a torus. All four of the original vertices now coincide, and we are left with two edges.

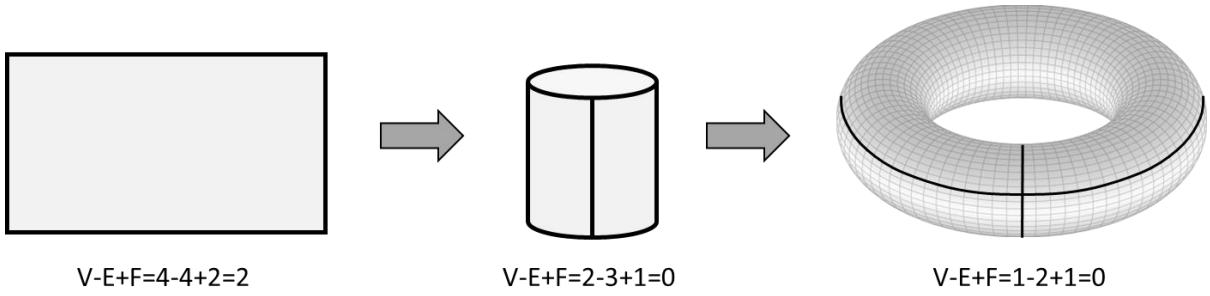


Figure 36. Euler characteristic for a rectangle, cylinder and torus

We can create additional surfaces by "glueing" the edges of a rectangle, as shown in Figure 37. The cylinder and torus from Figure 36 are repeated in Figure 37 but some additional notation, i.e., vertex labels and directional arrow for how the edges are to be joined. For example, to create the cylinder, we attach the two edges labelled with a which leads to vertex v_2 being identified with v_1 and vertex v_3 being identified with v_4. For the torus, we first attach the two edges labeled a and then the two edges labeled b. In the process, the four vertices are joined into one. The other three examples are more complex, involving a twist before joining two edges.

The rectangles, with the associated notation, in Figure 37 are known as fundamental polygons [47].

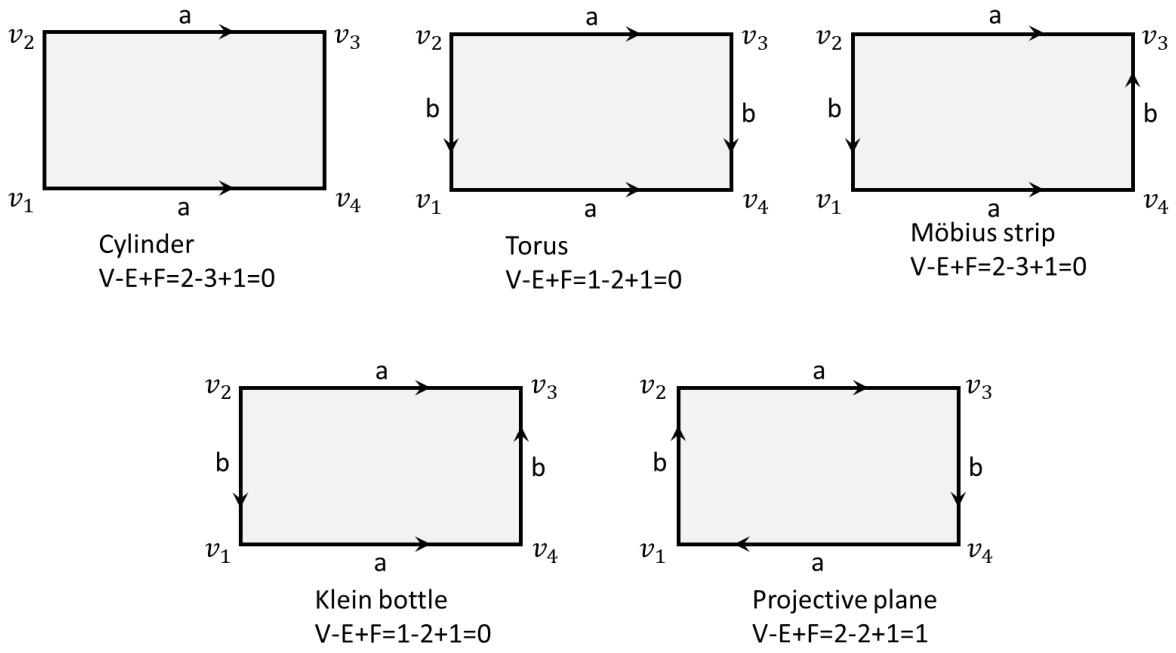


Figure 37. Surface created by "glueing" the edges of a rectangle

The result of folding and glueing a rectangle into a **Möbius strip** is shown in Figure 38. Vertices v_2 and v_4 coincide as do vertices v_1 and v_3, leaving us with two vertices. There are three edges (assuming the given partitioning of the Möbius strip), i.e., the solid vertical line, the dotted line and the dashed line. Clearly, there is only one face. So, we have $V - E + F = 2 - 3 + 1 = 0$. Alternately, if we remove the edge b (effectively removing the partition), there are no vertices, one edge and one face. This gives us $V - E + F = 0 - 1 + 1 = 0$.

From the Wikipedia article about the Möbius strip [48]:

> A Möbius strip, Möbius band, or Möbius loop is a surface that can be formed by attaching the ends of a strip of paper together with a half-twist. As a mathematical object, it was discovered by Johann Benedict Listing and August Ferdinand Möbius in 1858, but it had already appeared in Roman mosaics from the third century CE. The Möbius strip is a non-orientable surface, meaning that within it one cannot consistently distinguish clockwise from counterclockwise turns. Every non-orientable surface contains a Möbius strip.

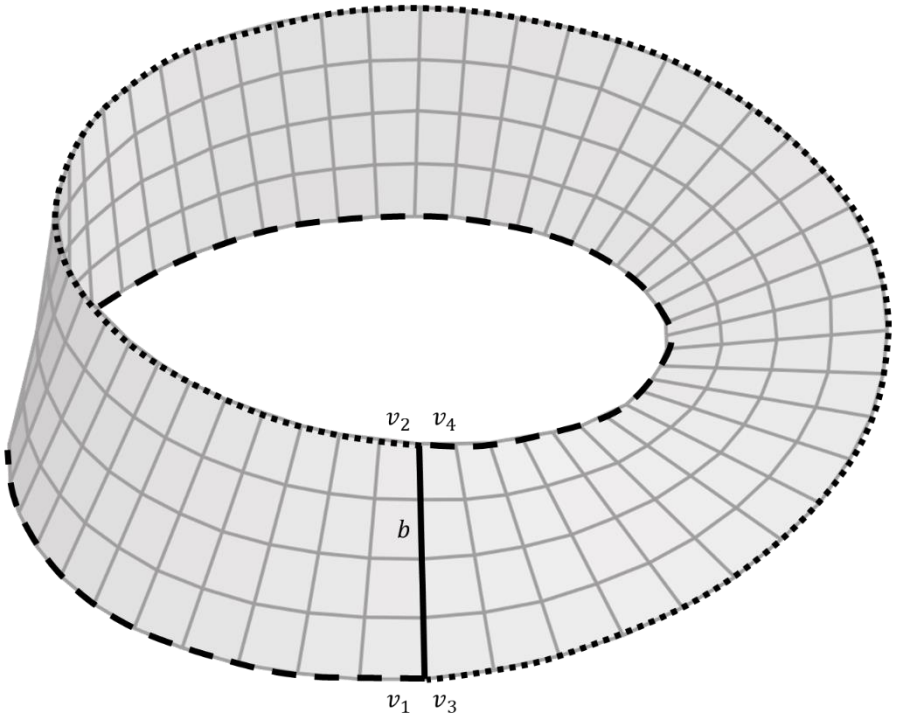


Figure 38. Möbius strip

So far, all the surfaces that we have encountered can be embedded in 3-dimensional space (i.e., $\mathbb{R}^3$). However, our next example (the Klein bottle) cannot be embedded in $\mathbb{R}^3$. When determining the Euler characteristic value of the Klein bottle, we need to be very careful when glueing the edges and vertices together. The rectangle is the only face. The two edges labelled a are glued together, as are the edges labelled b (but with a twist, as indicated by the arrow going in the opposite direction). This leaves us with 2 edges. In the last step, all four of the vertices are glued together, leaving just one vertex. Thus, $V - E + F = 1 - 2 + 1 = 0$.

The Klein bottle does **not** intersect itself. However, the typical visual representations in the literature show the Klein bottle as intersecting itself, e.g., Figure 39 is taken from the Wikipedia article on the Klein bottle [49]. The self-intersection issue can be resolved by placing the Klein bottle in 4 dimensions where the 4th dimension is time. In this way, the apparent points of intersection exist at different points in time and thus, do not intersect.

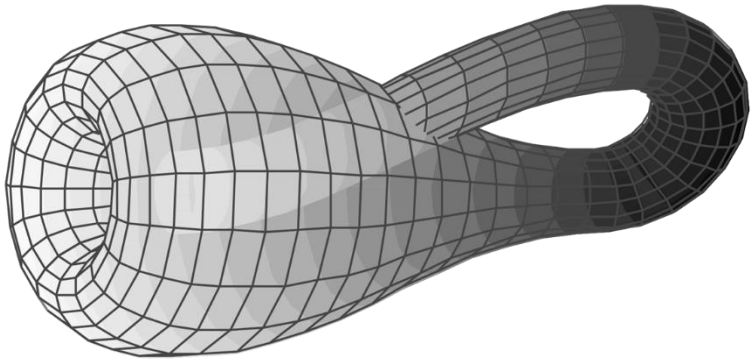


Figure 39. Klein bottle

A construction of the Klein bottle can be found in the Wikipedia article "Klein bottle" [49]. An animated version of the construction of the Klein bottle is provided in the YouTube video "What does the 4D Klein Bottle look like? [50].

In terms of the Euler characteristic, the real projective plane (often represented as $\mathbb{R}P^2$) has one face, 2 edges and 2 vertices (with v_1 glued to v_3, and v_2 glued to v_4). So, the Euler characteristic value for the real projective plane is $V - E + F = 2 - 2 + 1 = 1$. The mapping among the vertices is hard to see when using the fundamental polygon representation of the projective plane. However, it can be shown that the projective plane is homeomorphic to the unit disk with opposite points (aka antipodal point) identified with each other. Figure 40 shows an alternative representation of the projective plane where opposite points are equated (i.e., are the same point). For example, the two appearances of A in the figure are one and the same point.

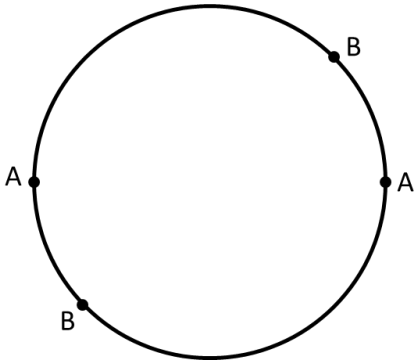


Figure 40. Alternative representation of projective plane

Although far from obvious, the fundamental polygon representation of the real project plane in Figure 36 is equivalent to 2-dimensional space $\mathbb{R}^2$ with the inclusion of points at infinity (aka ideal points). In the real projective plane, parallel lines intersect at one point (i.e., a point at infinity). Figure 41 depicts the point of infinity for one set of parallel lines. It is the same point at infinity in both the negative and positive direction, i.e., the infinite set of parallel lines intersect in one point. Further, for each different infinite set of parallel lines there is a different point at infinity.

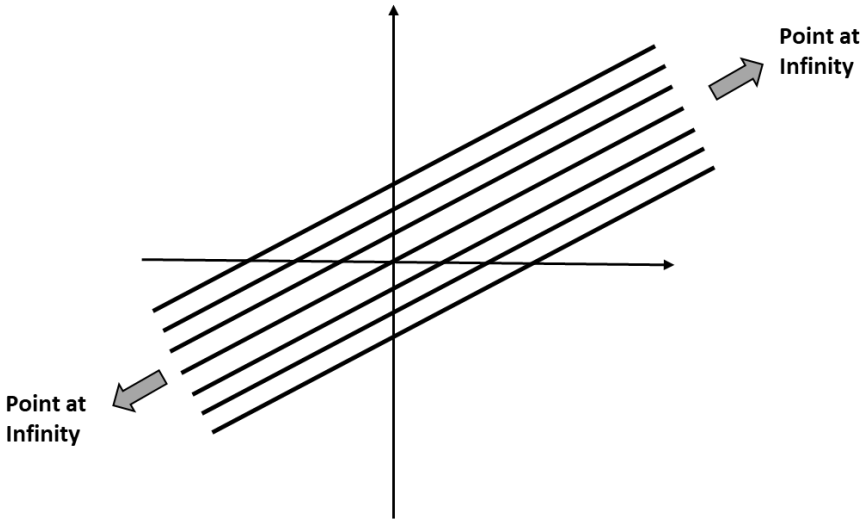


Figure 41. Point at Infinity for a set of parallel lines

The above representation is equivalent to the modified hemisphere model that we developed for single elliptic geometry (see Figure 27). The modified hemisphere model is, in turn, equivalent to the fundamental polygon for the real project plane (bottom right of Figure 36) and to the disk model in Figure 40. The various equivalences are explained beautifully in the YouTube video "M435 Ep 3 of 8 The Projective Plane RP2 Topology" [51]. The description in Section 3.3 of the Open University course on surfaces [54] is also very good.

3.3.3 Orientable Surfaces

As we saw in the previous section, different (non-homeomorphic) topological objects can have the same value for their Euler characteristic, e.g., the cylinder, the torus, Klein's bottle and the Möbius strip all have Euler characteristic of value 0 but are different topological objects. However, there are other topological invariants that can be used to distinguish among different topological objects. Orientability is one such invariant.

From the Wikipedia article on orientability [53]:

> A surface S in the Euclidean space $\mathbb{R}^3$ is orientable if a chiral two-dimensional figure *[i.e., a figure not identical to its mirror image]* **cannot** be moved around the surface and back to where it started so that it looks like its own mirror image. Otherwise, the surface is non-orientable. An abstract surface (i.e., a two-dimensional manifold) is orientable if a consistent concept of clockwise rotation can be defined on the surface in a continuous manner. That is to say that a loop going around one way on the surface can never be continuously deformed (without overlapping itself) to a loop going around the opposite way. This turns out to be equivalent to the question of whether the surface contains no subset that is homeomorphic to the Möbius strip. Thus, for surfaces, the Möbius strip may be considered the source of all non-orientability.

The last two sentences in the above quote are summarized in the following theorem.

Theorem 57. A surface is non-orientable if and only if it contains a Möbius strip.

Some textbooks on topology define orientability based on whether or not a surface has an embedded Möbius strip. For example, the following definition is taken from “Introduction to Topology Pure and Applied” [52].

> A surface that contains an embedded Möbius band is called non-orientable. A surface that does not contain an embedded Möbius band is called orientable.

The Möbius strip is a non-orientability surface. As can be seen in Figure 42, the directional arrow on the loop changes from clockwise to counterclockwise as the loop goes around the strip, which is a consequence of a surface not being orientable. The rule is “if a chiral figure can be returned to a given starting position but as its mirror image, then the surface is non-orientable.”

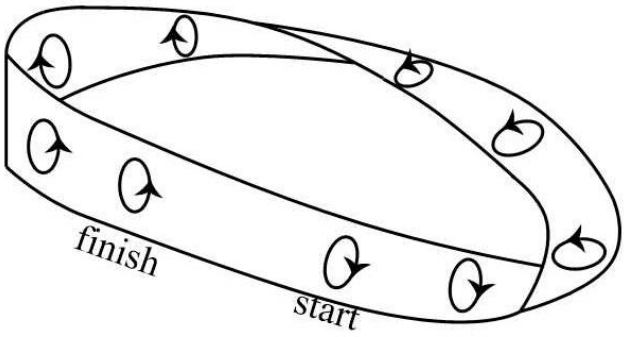


Figure 42. The Möbius strip is non-orientable

A strip with two half-twists (as shown in the upper left of Figure 43) is orientable since there is no way to move a chiral figure about the strip and return to its starting position in the form of its mirror image.

As shown in Figure 43, the strip with two half-twists can be cut, untwisted and the two edges can be rejoined (aligning exactly as before). The result is a cylinder, and in fact, a strip with two half-twist is homeomorphic to a cylinder.

[**Author’s Remark**: Tearing and reassembly, while preserving alignment of points along the cut, is allowed when determining whether two surfaces are homeomorphic. Most general, high-level, descriptions of homeomorphic surfaces gloss over this point, and yes, I am guilty of this in my open remarks.]

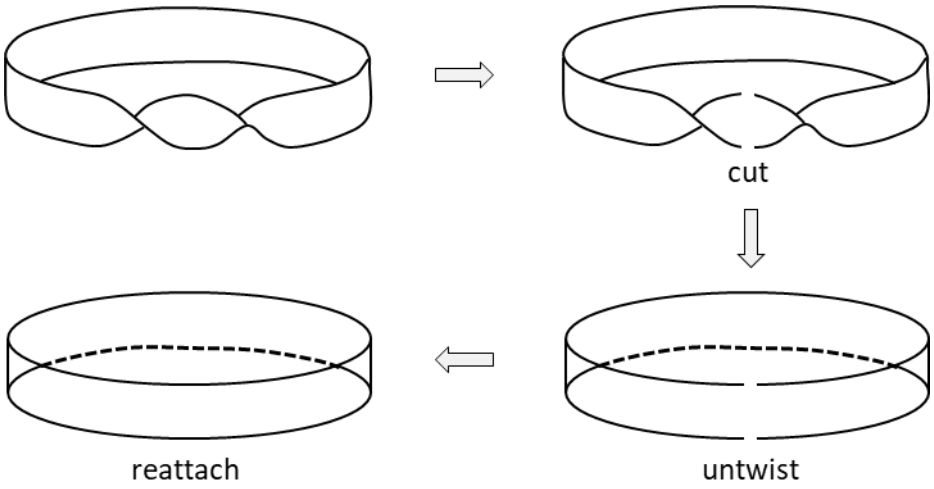


Figure 43. Strip with two twists

In general, an even number of half-twists in series will cancel each other out, and an odd number of half-twists in series will reduce to a single half-twist.

Figure 44 depicts several half-twists in parallel. If one traverses any one of the several possible loops on the surface with a chiral object (e.g., the loop shown as a dashed line), the chiral object will return to its initial position in the form of its mirror image. Thus, the surface is non-orientable.

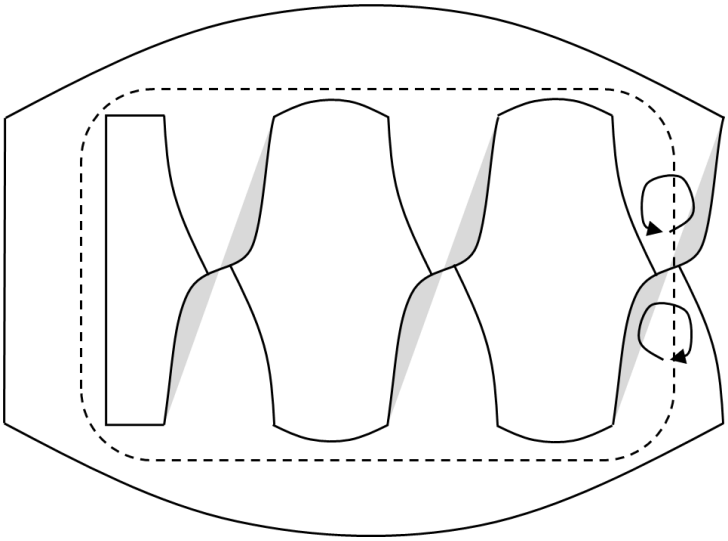


Figure 44. Half-twists in parallel

On the other hand, the surface in Figure 45 is orientable since the two half-twists cancel each other.

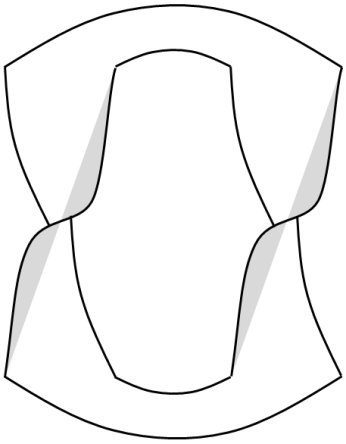


Figure 45. Orientable surface with two half-twists

By Theorem 57, a surface is non-orientable if it contains a Möbius band. To show that the Klein bottle and the projective plane are non-orientable, consider their representations in terms of fundamental polygons, as shown in Figure 46. When we identify (glue) the edges labelled b, the shaded region becomes a Möbius strip. This shows that the Klein bottle and the projective plane each contain a Möbius strip, and thus, are non-orientable.

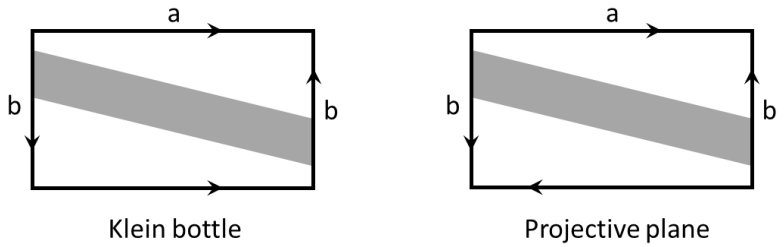


Figure 46. Klein and project plane are non-orientable

3.3.4 Boundary Numbers

The boundary of a surface is a concept that refers to the set of points that mark the edge or limit of a surface. A surface can have 0 or more boundaries. The number of boundaries of a surface is a topological invariant. For example, a cylinder has 2 boundaries (the circles at its top and bottom), the sphere, torus and Klein bottle have no boundaries, and the Möbius strip has just one boundary. In Section 3.5, a more precise definition of "boundary" will be provided.

Figure 47 depicts a surface with three boundaries. Three of the infinite number of boundary points (i.e., points on the boundary) are highlighted in the figure.

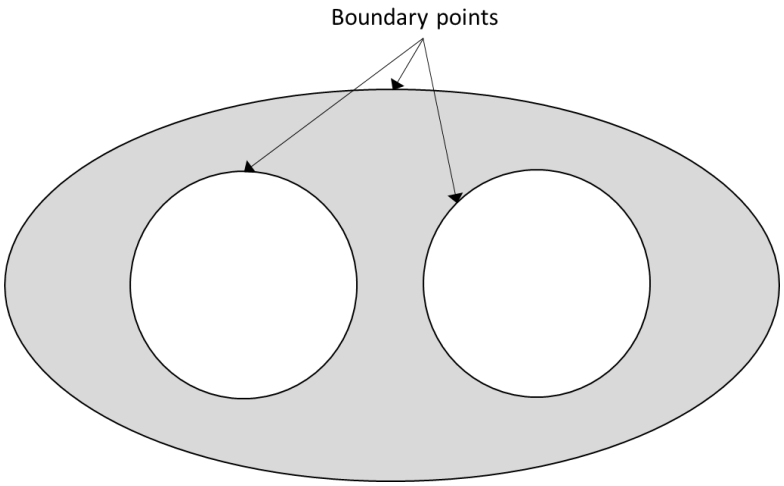


Figure 47. Surface with three boundaries

3.3.5 Subdivisions

A surface can be subdivided by adding vertices and edges in such a way as to preserve its Euler characteristic. On the left of Figure 48, two edges and 2 faces are added to a pentagon, resulting in the same Euler characteristic value of 2. Don't forget that for planar surfaces, we count the exterior of the surface as a face.

On the right of Figure 48, we have added 2 vertices, 3 edges and one face to a cube without changing the value of the Euler characteristic.

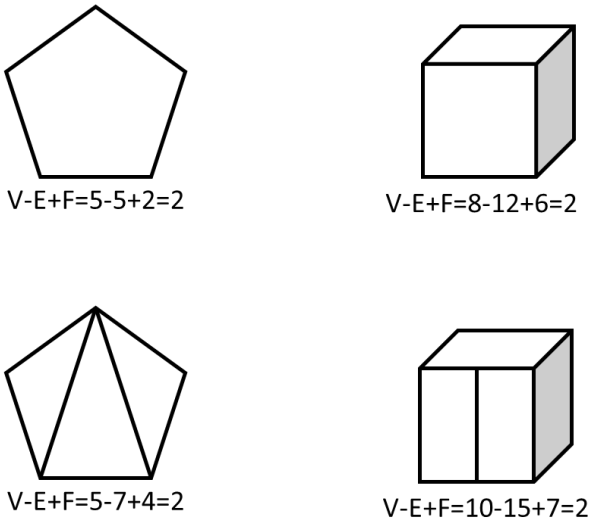


Figure 48. Subdividing a pentagon and a cube

There are various methods for subdividing a surface. The simplest method is called triangulation and is defined as follows:

- Each edge of a triangle is either on the boundary of the surface or shared with exactly one other triangle.
- The union of the triangles covers the entire surface.
- The intersection of any two triangles is either empty, a common vertex, or a common edge.

In the above definition the "triangles" are not necessarily flat unless a given face of a surface is also flat (in a plane). As used here, the term "triangle" is meant to indicate a three-sided surface that coincides with part of a larger surface.

Subdivision of a surface is not limited to triangles; any connected graph can be used. The concept is further detailed in the Wikipedia article Subdivision surface [55].

The Euler characteristic is a topological property, meaning it describes the overall shape of the surface regardless of how it's bent or stretched (as long as it doesn't develop holes or tears). Subdividing a surface simply cuts it up into smaller pieces, but doesn't change its fundamental shape.

While proving this result is mathematically involved, it essentially boils down to the fact that the Euler characteristic is calculated based on the number of vertices, edges, and faces in a subdivision. Regardless of how you subdivide the surface, the overall relationship between these elements will remain the same. We record this result in the following theorem.

Theorem 58. For any given surface, all of its subdivisions have the same Euler characteristic.

Proof: The proof involves concepts from algebraic topology (a topic not covered in this book). The interested reader is referred to Theorem 2.44 in the book by Hatcher [56]. ■

We can use Theorem 58 to compute the value of the Euler characteristic of a sphere. In the figure, the Euler characteristic of a sphere is computed using three different subdivisions, all with the same result of 2. On the left, the subdivision has only two faces, marked as F_1 (inside the triangle) and F_2 (outside the triangle). The middle subdivision has four faces, and the subdivision on the right has 3 faces. The edge between the two triangles in the surface on the right is necessary since the definition of a subdivision requires that the associated graph be connected. If we removed the edge, we would get $V - E + F = 6 - 6 + 3 = 3$ for the value of the Euler characteristic, which is incorrect for a sphere.

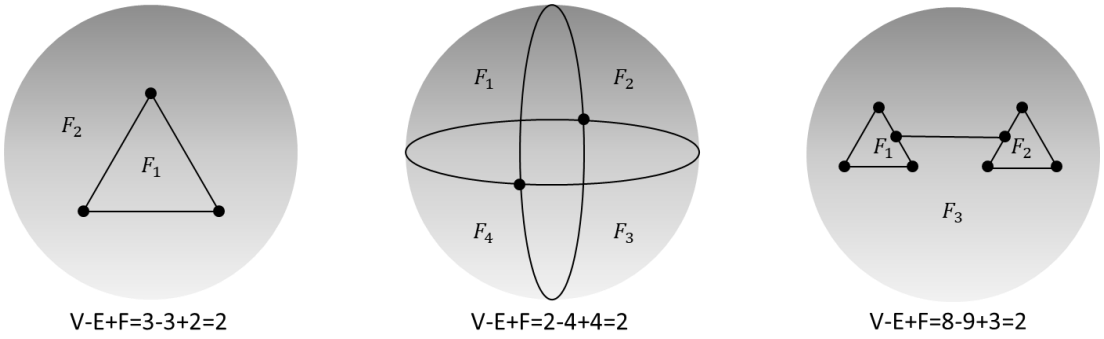


Figure 49. Euler characteristic of a sphere

3.3.6 Holes in Surfaces

The Euler characteristic of a point (and therefore also that of a closed or open disk) is 1.

Consider a surface S with Euler characteristic x. Create a subdivision of S which contains a disk. If we remove the face consisting of the disk from S, leaving the vertices and edges on the boundary of the disk with S, then the resulting surface has Euler characteristic $x - 1$. In general, we have the following theorem.

Theorem 59. Removing a disk (or surface homeomorphic to a disk, e.g., a non-intersecting polygon) from a surface reduces the Euler characteristic of the surface by 1.

3.3.7 Connected Sums

If a disk (or surface homeomorphic to a disk) is cut from each of two surfaces and the two surfaces are then joined along the two removed disks, the result is known as a **connected sum**. If the two surfaces being joined are S_1 and S_2, then the notation for their connected sum is $S_1 \# S_2$.

At the top of Figure 50, we have two tori (with a triangular-shaped face removed from each). Each torus now has one boundary at the removed triangular face. At the bottom of the figure, the two surfaces are joined along their triangular boundaries. The resulting surface has two holes and no boundary. It is known as a 2-hole torus. We could continue the process and an create n-holed tori for $n = 2,3,4, \ldots$

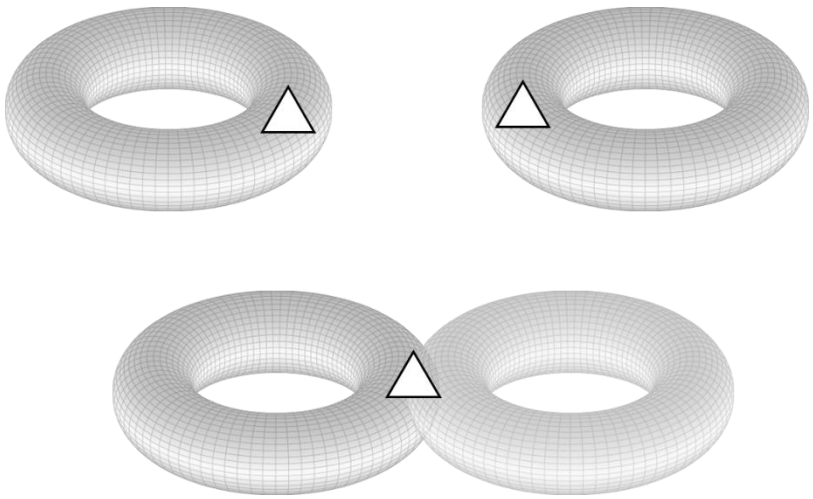


Figure 50. Connected sum of two tori

The following theorem allows us to compute the Euler characteristic of the connected sum of two surfaces.

Theorem 60. The Euler characteristic for the connected sum of surfaces S_1 and S_2 is the sum of the two Euler characteristics minus 2.

Proof: Assume the original subdivision of S_1 has v_1 vertices, e_1 edges and f_1 faces (this includes the v vertices and e edges of the face to be removed). Similarly, assume that S_2 initially has v_1 vertices, e_1 edges and f_1 faces (this includes the v vertices and e edges of the face to be removed). Further, assume that $e = v$ concerning the faces to be removed.

When the two surface are connected, the $2v$ vertices on the two removed faces are joined, leaving v vertices. Similarly, the $2e$ edges are joined, leaving e edges. The two faces are gone after the merger. So, the number of vertices, edges and faces for the connected sum is as follows:

$$V = v_1 + v_2 - v$$
$$E = e_1 + e_2 - e$$
$$F = f_1 + f_2 - 2$$

Thus, the Euler characteristic of the connected sum is

$$\begin{aligned} V - E + F &= (v_1 + v_2 - v) - (e_1 + e_2 - e) + (f_1 + f_2 - 2) \\ &= (v_1 - e_1 + f_1) + (v_2 - e_2 + f_2) - (v - e) \\ &= (v_1 - e_1 + f_1) + (v_2 - e_2 + f_2) - 2 \end{aligned}$$

The last equality in the above follows since $v = e$. ■

By Theorem 60, an n-holed torus has Euler characteristic $2 - 2n$. An n-holed torus is represented by the notation $\mathbb{T}^{\#n}$.

The number of holes in a surface is called the **genus** of the surface.

The connected sum of two spheres is again a sphere. So, this is not an interesting case.

On the other hand, the connected sum of two projective planes has Euler characteristic $1 + 1 - 2 = 0$, and the connected sum of n projective planes is $2 - n$. The connected sum of n projective planes is represented by the notation $\mathbb{P}^{\#n}$.

3.3.8 Classification of Surfaces

A **closed surface** contains a volume of space, enclosed from all directions; It consists of one connected, hollow piece that has no holes and doesn't intersect itself. A closed surface is one without a boundary, such as a torus or sphere, but not the cylinder. The following theorem classifies all closed surfaces.

[**Warning**: When we get to metric spaces and topological spaces in the following sections, the term "closed" is used differently. Perhaps a better term here would be "enclosed" but that is not commonly used, unfortunately.]

Theorem 61. Closed surfaces are classified as follows:

- Each two-sided, orientable closed surface is homeomorphic to $\mathbb{T}^{\#n}$ for some $n \geq 0$, n an integer. The case of $n = 0$ (i.e., a torus with no holes) is a sphere.
- Each one-sided, non-orientable closed surface is homeomorphic to $\mathbb{P}^{\#n}$ for some $n \geq 1$, n an integer.

The above theorem is taken from the topology book by Earl [57]. Earl goes on to make the following statement concerning the second part of Theorem 61:

> Making a connected sum with $\mathbb{P}$ is equivalent to sewing a Möbius strip into the surface. $\mathbb{P}$ itself can be made by introducing a Möbius strip into a sphere; to do this we might make a tear in the sphere and then, rather than gluing the tear back together, we could instead assign reverse arrows to the two sides of the tear, thus introducing a Möbius strip. So, the surface $\mathbb{P}^{\#k}$ can be thought of as a sphere with k Möbius strips sewed in.

The Klein bottle is homeomorphic to $\mathbb{P}\#\mathbb{P}$.

This is a good point to stop concerning surfaces. For a more detailed, yet basic, presentation on the classification of surfaces, see the unpublished paper “Classification of Surfaces” [58].

In the following section, we introduce a prelude to the more general topic of topological spaces.

3.4 Metric Spaces

3.4.1 Introduction

A **metric space** is a set of elements (aka points) along with a concept of distance between its elements. The distance is measured by a function called a metric or distance function. Many of the properties associated with metric spaces can be developed without the concept of distance in the more general context of what are known as topological spaces (to be covered in Section 3.5).

For example, the Euclidean spaces $\mathbb{R}^n$ for $n = 1,2,3, \ldots$ are metric spaces with distance function

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

where $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$. We will refer to this as the **Euclidean metric for $\mathbb{R}^n$**.

For $n = 1$, the distance function becomes

$$d(x,y) = \sqrt{(x-y)^2} = |x - y|$$

A metric space (M, d) consists of a set M and a distance function d that maps two elements of M to $\mathbb{R}$, i.e., $d: M \times M \to \mathbb{R}$, such that

- $d(x,x) = 0$, i.e., the distance between an element and itself is 0.
- For $x \neq y$, $d(x,y) \geq 0$.
- (Symmetry) $d(x,y) = d(y,x)$ for any $x, y \in M$.
- (Triangle Inequality) $d(x,y) \leq d(x,z) + d(z,y)$ for any $x, y, z \in M$.

A subtle point is that the mapping of each two elements of M to a real number implies that the distance between any two points in a metric space is finite.

If (X, d) is a metric space and Y is a subset of X, then the restriction of the metric d to $Y \times Y$ (call it d') is a metric on Y. The metric space (Y, d') is referred to as a **subspace** of (X, d).

3.4.2 Examples

The **Chebyshev distance** [59] between elements $x, y \in \mathbb{R}^n$, with standard coordinates x_i and y_i, respectively, is given by

$$d(x,y) = \max_{i=1,2,\ldots n} |x_i - y_i|$$

The set of elements in $\mathbb{R}^n$ along with the Chebyshev distance is a metric space.

. . .

Another distance function that defines a metric space over $\mathbb{R}^n$ is the taxicab distance [60], i.e.,

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

. . .

Take any set M and elements $x, y \in M$. The **discrete metric** for M is defined by

$$d(x,y) = \begin{cases} 1, & x \neq y \\ 0, & x = y \end{cases}$$

. . .

Consider the set of strings of equal length. For example, the characters in the string could be the digits 0 through 9 and the letters of the alphabet. The Hamming distance h counts the number of positions where the corresponding characters differ between two strings [61]. If we take the strings $x = dkmvk409489m0$ and $y = dkmvi709489t3$, then $h(x,y) = 4$ since the strings differ in 4 positions. The set of strings of a given length n with the Hamming distance form a metric space.

. . .

Let $C[a,b]$ be the set of continuous real-valued functions on the closed interval $[a,b]$ with any functions $f, g \in C[a,b]$. With the following metric, $C[a,b]$ is a metric space

$$d(f,g) = \int_a^b |f(x) - g(x)|\, dx$$

More generally, the following metric also gives us a metric space over $C[a,b]$

$$d_p(f,g) = \left(\int_a^b |f(x) - g(x)|^p \, dx \right)^{\frac{1}{p}}, \qquad p \geq 1$$

. . .

Let $\mathbb{R}^\infty$ be the set of all infinite sequences $\{x_i\}$. With the following metric, $\mathbb{R}^\infty$ is a metric space

$$d(x,y) = \sup_{i \in \mathbb{N}} |x_i - y_i|$$

The notation "$i \in \mathbb{N}$" means all values of i over the natural number, i.e, $i = 1,2,3, \ldots$

The term “sup” is short for “supremum” where the supremum [62] of a subset S of a partially ordered set P (e.g., the real numbers) is the least element in P that is greater than or equal to every element in S if such an element exists. For example, the sup of $S = \left\{1 - \frac{1}{n}\right\}$ for $n = 1,2,3, \ldots$ is 1 whereas the maximum of S does not exist as an element of S.

. . .

Let S be any set and let $B(S)$ denote the set of bounded real-valued functions on S with metric

$$d(f,g) = \sup\{|f(s) - g(s) : s \in S|\}$$

As defined, $B(S)$ with metric d is a metric space.

. . .

A distance function known as the Cayley–Klein metric can be applied to the Poincaré disk model for hyperbolic geometry [25].

Given two distinct points p and q inside the disk, the unique hyperbolic line connecting them intersects the boundary at two ideal points, a and b. Label them so that the points are, in order,

a, p, q, b, so that $d_2(a,q) > d_2(a,p)$ and $d_2(p,b) > d_2(q,b)$ where d_2 is the Euclidean distance as defined in $\mathbb{R}^2$, see Figure 51. The Cayley-Klein metric is defined as follows:

$$d(p,q) = \ln\left[\frac{d_2(a,q)\cdot d_2(p,b)}{d_2(a,p)\cdot d_2(q,b)}\right]$$

where ln is the log base e, i.e., the natural log.

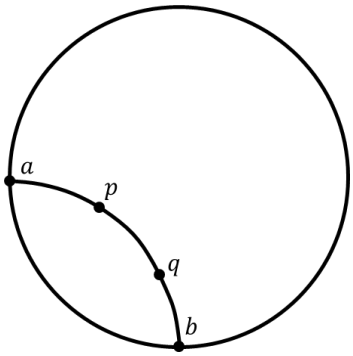


Figure 51. Cayley–Klein metric

3.4.3 Open and Closed Sets

On the real line (i.e., $\mathbb{R}$), $(0,1)$ and $(-3,4)$ are examples of open sets, and $[-1,7]$ and $[5,13]$ are examples of closed sets. In the real plane (i.e., $\mathbb{R}^2$), the interior of a circle is an example of an open set, and a disk (circle and its interior) is an example of a closed set. The concepts of open and closed sets are generalized in the context of metric spaces.

We start by defining an **open ball** in a metric space (M, d) as the set

$$B(x;r) = \{y \in X : d(x,y) < r\}$$

$B(x{:}\, r)$ is said to be the open ball centered at x with radius r.

In $\mathbb{R}^3$ with the Euclidean metric, $B\big((0,0);5\big)$ is the interior of a sphere centered at the origin with radius 5.

In the metric space of continuous function on the interval $[-2,2]$, the open ball with center $f(x) = x^3 - 2x$ and with radius 3 is shown in Figure 52. In words, the open ball includes only continuous functions on the interval $[-2,2]$ which are within (less than) 3 from $f(x) = x^3 - 2x$ for each value of x on the interval (dark gray area in the figure).

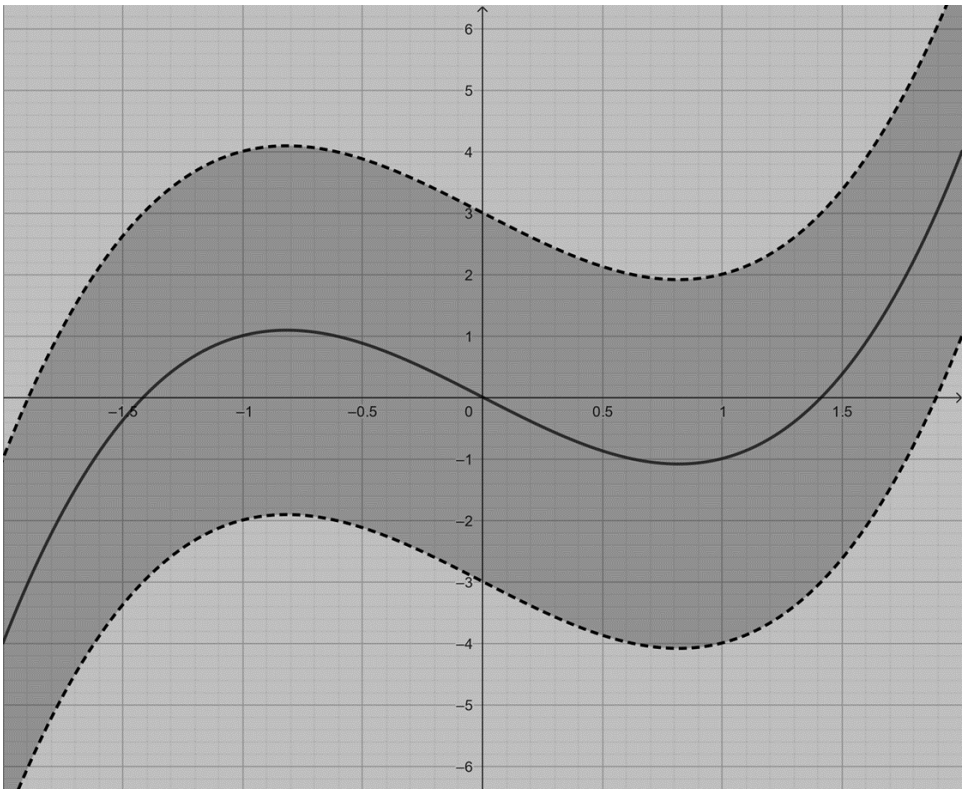


Figure 52. Open ball about f(x) = x^3 -2x

Since the distance between any two points in a metric space M is a real number (i.e., finite), we have that the union of all open balls about a given point $x \in M$ equals the entire metric space, i.e.,

$$\bigcup_{r>0} B(x;r) = M$$

Since $d(x,y) = 0$ if and only if $x = y$, we have that

$$\bigcap_{r>0} B(x:r) = \{x\}$$

Let Y be a subset of X (which we write as $Y \subseteq X$). A point $x \in X$ is defined to be an **interior point** of Y if there exists $r > 0$ such that $B(x;r) \subset Y$. The set of interior points of Y is referred to as the **interior** of Y, and it is represented by $int(Y)$. Clearly, every interior point of Y is an element of Y and so, $int(Y) \subseteq Y$.

As a reminder from set theory, $A \subset B$ means A is a proper subset of B (i.e., A cannot be all of B), and $A \subseteq B$ means A is a subset of B and could be equal to B.

A subset Y of X is defined to be **open** if every point of Y is an interior point of Y, i.e., if $int(Y) = Y$. In all cases, the empty set $\emptyset$ and the entire space X are open subsets of X.

The next several theorems may seem obvious, but they need to be proved.

Theorem 62. An open ball of a metric space X is an open subset of X.

Proof: Take any open ball $B(x;r)$ in metric space (X,d). We need to show every point $y \in B(x;r)$ is an interior point of $B(x;r)$.

Since $y \in B(x;r)$, $s = r - d(x,y) > 0$.

Next, take any $z \in B(y;s)$ which implies $d(y,z) < s$. Using the triangle property of metric spaces, we have

$$d(x,z) \leq d(x,y) + d(y,z) < d(x,y) + s = r$$

Thus, $z \in B(x;r)$ which implies $B(y;s) \subset B(x;r)$. So, y in an interior point of $B(x;r)$. ■

Theorem 63. The interior of subset Y of a metric space (X,d) is an open set, i.e., $int(int(Y)) = int(Y)$.

Proof: To show that $int(Y)$ is open, we need to show every point $y \in int(Y)$ is an interior point of $int(Y)$. Since $y \in int(Y)$, there exists $r > 0$ such that $B(y;r) \subset Y$. If we can show that $B(y;r) \subset int(Y)$, we are done.

Since $B(y;r)$ is open by Theorem 62, there exists for each $x \in B(y;r)$ an open ball $B(x;s) \subset B(y:r)$. However, $B(y;r) \subset Y$ and so, $B(x;s) \subset Y$. Thus, each $x \in B(y;r)$ is in $int(Y)$ which implies $B(y;r) \subset int(Y)$. ■

Theorem 64. Any union of open subsets in a metric space (X,d) is an open set.

Proof: Let $\{Y_a\}_{a \in A}$ be a collection of open sets in the given metric space, and let $Y = \bigcup_{a \in A} Y_a$. Take any $y \in Y$. It must be that $y \in Y_b$ for some $b \in A$. Since Y_b is, by assumption, an open set, there exists an open ball $B(y;r) \subset Y_b \subseteq Y$ which implies y is an interior point of Y. Thus, Y is open. ■

The above theorem works for infinite collections of open subsets, but the same is not true for infinite intersections. For example, consider the collection of open subsets $\left\{\left(0,1+\frac{1}{n}\right)\right\}$ for $n = 1,2,3,\ldots$ in the metric space $\mathbb{R}$ with the Euclidean metric. The intersection of these subsets is $(0,1]$ which is not open. However, the intersection of a finite number of open sets is open.

Theorem 65. The intersection of finite number of open subsets in a metric space (X,d) is an open set.

Proof: Let $\{Y_i\}, i = 1,2,\ldots,n$ be a collection of open subsets in a metric space and let $Y = \bigcap_{i=1}^{n} Y_i$. Take any $y \in Y$. Since for each of the Y_i, there exist radii r_i such that $B(y;r_i) \subset Y_i, i = 1,2,\ldots,n$. If we let $r = \min(r_1, r_2, \ldots, r_n)$, then $r > 0$ since the minimum of a finite set of positive numbers is a positive number. So, $B(y;r) \subset Y_i, i = 1,2,\ldots,n$ which implies $B(y;r) \subset Y$. Thus, Y is an open subset of (X,d). ■

Theorem 66. A subset of a metric space is open if and only if it is the union of open balls.

Proof: By Theorem 62 and Theorem 64, the union of a set of open balls is an open set.

Going in the other direction, assume Y is an open set. This implies that for each $y \in Y$ there exist an open ball $B(y;r_y) \subset Y$. Forming the union of all the open balls, we have that $Y = \bigcup_{y \in Y} B(y;r_y)$. ■

A set can be an open set in a subspace but not open in the containing space. For example, the unit disk $x^2 + y^2 < 1$ is open in $\mathbb{R}^2$ but not open in $\mathbb{R}^3$. In fact, the only open set in $\mathbb{R}^2$ that is open in $\mathbb{R}^3$ is the empty set. We do, however, have the following theorem.

Theorem 67. A set A is open in subspace Y of metric space X if and only if there exist an open set B in X such that $A = B \cap Y$.

Proof: See the YouTube video by The Math Sorcerer [63]. ■

We can illustrate Theorem 67 using our disk example. If we let A be the open unit disk centered at the origin, i.e., $A = \{(x, y): x^2 + y^2 < 1\} \subset \mathbb{R}^2$ and let B be the open unit sphere centered at the origin, i.e., $B = \{(x, y, z): x^2 + y^2 + z^2 < 1\} \subset \mathbb{R}^3$, then $A = B \cap \mathbb{R}^3$ which implies that A is open in $\mathbb{R}^2$.

We conclude our discussion of open sets with this important definition:

> Two metrics on a set X are said to be **equivalent** if they determine the same open subsets.

. . .

For a subset Y of a metric space (X, d). A point $x \in X$ is **adherent** to Y if for every $r > 0$,

$$B(x; r) \cap Y \neq \emptyset$$

The **closure** of Y, represented as $\overline{Y}$, is the set of all points in X that are adherent to Y. By definition, each point of Y is adherent to Y which implies $Y \subseteq \overline{Y}$.

Some examples,

- Each point in $A = \{(x, y): x^2 + y^2 < 1\} \subset \mathbb{R}^2$ is adherent to A (assuming the Euclidean metric on $\mathbb{R}^2$). All the point on the circle $x^2 + y^2 = 1$ are also adherent to A. Thus, $\bar{A} = \{(x, y): x^2 + y^2 \leq 1\}$.
- Each point in $B = \left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\right\} \subset \mathbb{R}$ is adherent to B (assuming the Euclidean metric on $\mathbb{R}$). In addition, 0 is adherent to B. Thus, $\overline{B} = B \cup \{0\}$.
- In $\mathbb{R}$, the closure of $C = [0,1) \cup (1,2]$ is $[1,2]$. The closer of $D = [0,1) \cup (2,3]$ is $[0,1] \cup [2,3]$.

A subset Y of metric space (X, d) is defined to be **closed** if $Y = \overline{Y}$. In all cases, the empty set Ø and the entire space X are closed subsets of metric space (X, d). In the above examples, A, B, C and D are not closed. Set A is open, but sets B, C and D are not open. So, open and closed are not opposite concepts.

. . .

The following concepts lead us to an alternate (but equivalent) definition of a closed set in a metric space.

A point x in metric space (X, d) is a **limit point** (or accumulation point) of a subset Y of X if every open ball with x at its center contains a point of Y other than x. The following results (stated without proof) hold true for limit points:

- x is a limit point of a subset Y of metric space (X, d) if and only if there exists an infinite sequence $\{x_1, x_2, x_3, \dots\}$ such that $d(x_i, x) \to 0$ as $i \to \infty$, and $x_i \neq x$ for all i. In this case, we say that the sequence **converges** to x. Symbolically, $\{x_n\} \to x$ if

$$\forall \epsilon > 0 \left(\exists N \in \mathbb{N} \left(\forall n \in \mathbb{N} \left(n \geq N \implies |x_n - x| < \epsilon\right)\right)\right)$$

- The set of limits points for a subset Y of metric space (X, d) is a closed set.
- Two metrics d_1, d_2 on a set X are equivalent if and only if the convergent sequences in (X, d_1) are the same as the convergent sequences in (X, d_2).

Exercise for the reader: Let (X, d) be a metric space. Define $d' : X \times X \to \mathbb{R}$ as $d'(x, y) = \min\big(1, d(x, y)\big)$. Show that d' is a metric on X and that it is equivalent to d. Thus, every metric is equivalent to a bounded metric.

A point x in metric space (X, d) is an **isolated point** of subset Y of X if there exist $r > 0$ such that $B(x; r) \cap Y = \{x\}$.

The following result (stated without proof) gives us another way to define the closure of a set, and thus, an alternate definition of a closed set.

> The closure of a subset Y in metric space (X, d) is the disjoint union of the limit points and isolated points of Y.

The above result highlights the fact that a closed set can have isolated points. For example, the set $\{(x, y) : x^2 + y^2 \leq 1\} \cup \{(2,3), (-5,4)\}$ is a closed subset of the metric space $\mathbb{R}^2$ with the Euclidean metric.

The **boundary of a subset** Y of a metric space (X, d) is defined to be the set of points adherent to both Y and the complement of Y (i.e., points in X but not in Y). The boundary of Y is denoted ∂Y. The following results (stated without proof) effectively provide alternate definitions of open and closed subsets of a metric space:

- A subset Y is open if and only if $Y \cap \partial Y$ is empty.
- A subset Y is closed if and only if $\partial Y \subseteq Y$.

Theorem 68. The limit of a convergent sequence in a metric space is unique.

Proof: Assume that the sequence $\{x_n\}$ converges to x and y. For each n, we have that

$$d(x, y) \leq (x, x_n) + d(x_n, y)$$

So, as $n \to \infty$, the two terms on the right of the above inequality go to 0 which implies $d(x, y) = 0$ and so, $x = y$. ■

Theorem 69. If Y is a subset of a metric space (X, d), then $x \in X$ is adherent to Y if and only if there is a sequence in Y that converges to x.

Proof: If there exists a sequence in Y that converges to x, then every open ball centered at x contains points of the sequence, so that x is adherent to Y.

Going in the other direction; if x is adherent to Y, then for each integer $n \geq 1$, then there exists some point $x_n \in B\left(x; \frac{1}{n}\right) \cap Y$. The sequence $\{x_n\}, n = 1,2,3, \ldots$ has the property that $d(x, x_n) < \frac{1}{n} \to 0$, as $n \to \infty$ and so, x_n converges to x. ■

. . .

As an example, consider set Y that includes the unit closed disk centered at the origin, with the points $(0,0)$ and $(1,0)$ removed, and the set $\left\{(2,0), \left(\frac{3}{2}, 0\right), \left(\frac{5}{4}, 0\right), \left(\frac{9}{8}, 0\right), \ldots, \left(\frac{2^i+1}{2^i}, 0\right), \ldots\right\}$ included. Assume the context is the metric space $\mathbb{R}^2$ with the Euclidean metric.

- The set of adherent points (i.e., the closure) is the closed unit disk, i.e., the set $\{(x, y): x^2 + y^2 \leq 1\}$. Since $Y \neq \overline{Y}$, Y is not closed.
- The interior of Y is the unit open disk minus the point $(0,0)$. So, $Y \neq int(Y)$, thus, Y is not open.
- The boundary of Y is the circle $x^2 + y^2 = 1$ and the point $(0,0)$.
- All the points in $\left\{(2,0), \left(\frac{3}{2}, 0\right), \left(\frac{5}{4}, 0\right), \left(\frac{7}{8}, 0\right), \ldots, \left(\frac{2^i+1}{2^i}, 0\right), \ldots\right\}$ are isolated points.
- The points $(0,0)$ and $(1,0)$ are limit points (as are all the other points in the closed unit disk).

. . .

Theorem 70. If Y is a subset of a metric space (X, d), then the closure of Y is closed, i.e., $\overline{Y} = \overline{\overline{Y}}$.

Proof: See the YouTube video by The Math Sorcerer [64]. ■

We need a few additional definitions and results from set theory before proceeding.

The **complement of a set** A, denoted by A^c, is the set of elements not in A relative to some universal set U of interest. For example, the complement of the interval $[1,3]$ in $\mathbb{R}$ is $(-\infty, 1) \cup (1, \infty)$.

The **relative complement** of A with respect to a set B, also known as the set difference of B and A, written $B \backslash A$ or sometimes as $B - A$, is the set of elements in B that are not in A. In terms of the universal set, we can write $A^c = U \backslash A$.

Theorem 71. The following set relationships hold true:

$$A\backslash\bigcap_{e\in E} B_e = \bigcup_{e\in E}(A\backslash B_e)$$

$$A\backslash\bigcup_{e\in E} B_e = \bigcap_{e\in E}(A\backslash B_e)$$

Proof: We prove the first equation using a series of "if and only if" statements. Proof of the second equation is similar and left to the reader.

$x \in A\backslash \bigcap_{e\in E} B_e$ if and only if $x \in A$ and $x \notin \bigcap_{e\in E} B_e$

if and only if $x \in A$ and $x \notin B_f$ for at least one $f \in E$

if and only if $x \in A\backslash B_f$ for at least one $f \in E$

if and only if $x \in \bigcup_{e\in E}(A\backslash B_e)$. ■

Theorem 72. A subset Y of a metric space (X, d) is closed if and only if the complement of Y is open.

Proof: See the YouTube video from The Math Sorcerer [65].

Theorem 73. The intersection of any collection of closed sets is closed. The union of any finite collection of closed sets is closed.

Proof: Assume the sets in question are subsets of the metric space (X, d). Let $\{Y_{a\in A}\}$ be a collection of closed sets. Applying Theorem 71, we have

$$X\backslash\bigcap_{a\in A} Y_a = \bigcup_{a\in A}(X\backslash Y_a)$$

Since each Y_a is closed, each $X\backslash Y_a$ is open (by Theorem 72). By Theorem 65, $\bigcup_{a\in A}(X\backslash Y_a)$ is open, and so, $X\backslash \bigcap_{a\in A} Y_a$ is also open. Applying Theorem 72, we have that $\bigcap_{a\in A} Y_a$ is open.

The proof of the second part of the theorem follows a similar approach, and is left as an exercise for the reader. ■

Consider the collection of closed sets in $\mathbb{R}$:

$$Y_n = \left[-\frac{1}{n}, 2-\frac{1}{n}\right], n = 1,2,3,\ldots$$

$$Z_n = \left[\frac{1}{n}, \frac{2^{n+1}-1}{2^n}\right], n = 1,2,3,\ldots$$

The intersection of all the Y_n is closed, i.e.,

$$\bigcap_{i=1}^{\infty}\left[-\frac{1}{n}, 2-\frac{1}{n}\right] = [0,1]$$

but the union of the Z_n is open, i.e.,

$$\bigcup_{i=1}^{\infty}\left[\frac{1}{n},\frac{2^{n+1}-1}{2^n}\right]=(0,2)$$

. . .

Theorem 74. The complement of a one-point set in a metric space (X,d), consisting of more than one point, is open.

Proof: Let $\{p\}$ be the one-point set in question. We want to show the $X\backslash\{p\}$ is open.

Choose any point $x\in X\backslash\{p\}$. Since $x\neq p$ (which is true since we assumed X has more than one point), $d(x,p)>0$.

Let $r=\frac{d(x,p)}{2}>0$ and consider the open ball $B(x;r)$. If we can show $B(x;r)\subset X\backslash\{p\}$, then we are done.

For any point $y\in B(x;r)$, we have $d(x,y)<r=\frac{d(x,p)}{2}$. Using the triangle inequality, we have

$$d(y,p)\geq d(x,p)-d(x,y)$$

and since $d(x,y)<\frac{d(x,p)}{2}$, we have

$$d(y,p)>d(x,p)-\frac{d(x,p)}{2}=\frac{d(x,p)}{2}>0$$

Since $d(y,p)>0$, it must be that $y\neq p$, and so, $y\in X\backslash\{p\}$. Thus, $B(x;r)\subset X\backslash\{p\}$. ■

3.4.4 Completeness

Consider the metric space $(0,1)$ with the Euclidean metric. The adjacent points in the sequence $\left\{\frac{1}{n}\right\},n=1,2,3,\ldots$ become arbitrarily close as $n\to\infty$ but they do not converge in the metric space since we have excluded the point 0. In some sense, this metric space is incomplete.

Another example of an incomplete metric space is the open unit disk, i.e., $\{(x,y):x^2+y^2<1\}$ with the Euclidean metric. There are many sequences in the open unit disk whose points become arbitrarily close but do not converge within the metric space, e.g., $\left\{\left(0,\frac{n}{n+1}\right)\right\},n=1,2,3,\ldots$

The sequences in the previous examples are known as Cauchy sequences [66]. In general, the sequence $\{x_n\},n=1,2,3,\ldots$ in the metric space (X,d) is a **Cauchy sequence** if

$$\lim_{m,n\to\infty}d(x_m,x_n)=0$$

More precisely, $\{x_n\},n=1,2,3,\ldots$ is a Cauchy sequence if for any $\epsilon>0$, the exists positive integer N such that $d(x_m,x_n)<\epsilon$ whenever $m,n>N$.

There is a subtle point to be made here, i.e., it is not sufficient that $d(x_n,x_{n+1})\to\infty$ as $n\to\infty$ for a sequence to be a Cauchy sequence. Consider the sequence $\{\sqrt{n}\}$. We have that

$$x_{n+1} - x_n = \sqrt{n+1} - \sqrt{n} = \frac{1}{\sqrt{n+1} + \sqrt{n}} < \frac{1}{2\sqrt{n}}$$

which approaches 0 as $n \to \infty$. However, as n increases, the terms $\{\sqrt{n}\}$ become arbitrarily large. So, for any index n and distance r, there exists an index m sufficiently large such that a$\sqrt{m} - \sqrt{n} > r$. So, $\{\sqrt{n}\}$ is not a Cauchy sequence.

The utility of the Cauchy sequence concept lies in the fact that in a **complete metric space** (one where all Cauchy sequences converge to a limit in the metric space), the criterion for convergence depends only on the terms of the sequence itself, and that does not require knowing the limit of the sequence.

The real number line $\mathbb{R}$ and n-dimensional Euclidean space $\mathbb{R}^n$ are two common examples of complete metric spaces [67].

Theorem 75. A convergent sequence is a Cauchy sequence.

Proof: Let $\{x_n\}, n = 1,2,3, \ldots$ converge to x in metric space (X, d). By the triangle inequality for metric spaces, we have

$$d(x_m, x_n) \leq d(x_m, x) + d(x, x_n)$$

The right-hand side of the inequality approaches 0 as $m, n \to \infty$, and so, $\{x_n\}$ is a Cauchy sequence. ■

Theorem 76. If a subsequence of Cauchy sequence converges to x, then so does the entire Cauchy sequence.

Proof: Let (X, d) be a metric space, and $\{x_n\}$ be a Cauchy sequence with subsequence $\{x_{n_k}\}$ that converges to x. We want to show $x_n \to x$ as $n \to \infty$.

Since $\{x_n\}$ is a Cauchy sequence, for every $\epsilon > 0$, there exist a positive integer N such that for all $n, m > N$, the following holds true

$$d(x_n, x_m) < \frac{\epsilon}{2}$$

By hypothesis, there exists $L > 0$ such that for all $n_k > L$, the following holds true

$$d(x_{n_k}, x) < \frac{\epsilon}{2}$$

If we let $M = \max(N, L)$, then for all $n, m, n_k > M$, we have

$$d(x_n, x) \leq d(x_n, x_{n_k}) + d(x_{n_k}, x) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

Thus, $x_n \to x$ as $n \to \infty$. ■

The next two theorems relate the concepts of closed and complete.

Theorem 77. A closed subspace of a complete metric space is complete.

Proof: Let Y be a closed subspace of a complete metric space (X, d). In addition, let $\{y_n\}$ be a Cauchy sequence in Y and consequently, also a Cauchy sequence in X. Since X is complete, $y_n \to y$ for some $y \in X$. However, Y is closed, and so, $y \in Y$ (this follows from Theorem 69 and the definition of closed). So, all Cauchy sequences in Y converge to a point in Y, and thus, Y is complete. ■

Theorem 78. A complete subspace Y of a metric space (X, d) is closed in X.

Proof: For Y to be closed, it must be that $Y = \overline{Y}$ which, in turn, implies that Y must contain all its adherent points. Let point $x \in X$ be adherent to Y. This means for every $r > 0, B(x; r) \cap Y \neq \emptyset$. Take $r = \frac{1}{n}, n = 1,2,3, \ldots$ and take a representative $x_n \in B(x; 1/n) \cap Y$ for $n = 1,2,3, \ldots$ So, we have $x_n \to x$ as $n \to \infty$. By Theorem 75, $\{x_n\}$ is a Cauchy sequence, and since Y is complete, $\{x_n\}$ converges to some point $y \in Y$. The limit of a sequence is unique by Theorem 68, and so, we have $x = y \in Y$. Thus, Y is closed since it contains each of its adherent points. ■

. . .

A subset Y of a topological space X is said to be a **dense** subset of X if any of the following equivalent conditions are satisfied:

- The smallest closed subset of X containing Y is X.
- $\overline{Y} = X$
- $int(X \backslash Y) = \emptyset$
- Every point in X either belongs to Y or is a limit point of Y.
- For every $x \in X$, every open set U of X intersects Y, i.e., $U \cap Y \neq \emptyset$.
- Y intersects every non-empty open subset of X.

Theorem 79. The rational numbers $\mathbb{Q}$ are dense in the metric space of real numbers $\mathbb{R}$ under the Euclidean metric.

Proof: We first show that between any two real numbers there exists a rational number (i.e., number of the form p/q where p and q are integers).

Let x and y be real numbers. Assume $x < y$, and then consider $y - x$. If the difference is greater than 1 then we are done because there is an integer between them x and y. Otherwise, multiply $y - x$ by some large enough integer n so that $n(y - x) > 1$ which implies there exists an integer m such that $nx < m < ny$ and so, $x < \frac{m}{n} < y$.

Using the above result and taking $\epsilon > 0$, we have a rational number r_ϵ between a given real number x and $x + \epsilon$. Taking $\epsilon = 1, \frac{1}{2}, \frac{1}{3}, \ldots$ gives us a sequence of rational numbers $r_{\frac{1}{n}}$ that converge to x, and by Theorem 69, $x \in \overline{\mathbb{Q}}$. ■

The set $(0,1)$ is not dense in the metric space $\mathbb{R}$ but it is dense in the metric space $[0,1]$.

The set of integers in not dense anywhere in $\mathbb{R}$. In general, a subset Y of a metric space X is said to be **nowhere dense** in X if its closure has no interior points, i.e., $int(\overline{Y}) = \emptyset$.

The Baire category theorem gives sufficient conditions for a metric space to be a Baire space (a metric space such that the intersection of countably many dense open sets is still dense).

Theorem 80. (Baire category theorem) If $Y_i, i = 1, 2, 3, \ldots$ is a collection of dense open subsets in a complete metric space (X, d) then

$$\bigcap_{i=1}^{\infty} Y_i$$

is also dense in X.

Proof: The Wikipedia article "Baire category theorem" [69] states two more general forms of the theorem. The proof of BCT1 in the Wikipedia article can be applied to metric spaces. ■

A subset Y of a topological space is said to be **nowhere dense** or rare if its closure has empty interior, i.e., $int(\overline{Y}) = \emptyset$.

Theorem 81. A subset Y of a metric space X is nowhere dense in X if and only if $X \backslash \overline{Y}$ is a dense open subset of X.

Proof: See the question "Prove that if a set is nowhere dense iff the complement of the closure of the set is dense" from Mathematics Stack Exchange [82]. ■

By taking the complements of the sets in the Baire Category Theorem, we get the following equivalent version.

Theorem 82. If $Y_i, i = 1, 2, 3, \ldots$ is a collection of nowhere dense subsets in a complete metric space (X, d) then

$$int\left(\bigcup_{i=1}^{\infty} Y_i\right) = \emptyset$$

3.4.5 Products of Metric Spaces

A product metric is a metric on the Cartesian product of finitely many metric spaces $(X_1, d_1), (X_2, d_2), \ldots, (X_n, d_n)$. This is represented as $X_1 \times X_2 \times \ldots \times X_n$. Various metrics will yield a metric space on the product, e.g., the **p-norms** [70] for fixed $p \in [1, \infty)$

$$d_p\big((x_1, \ldots, x_n), (y_1, \ldots, y_n)\big) = (d_1(x_1, y_1)^p + \cdots + d_n(x_n, y_n)^p)^{\frac{1}{p}}$$

$$d_\infty\big((x_1, \ldots, x_n), (y_1, \ldots, y_n)\big) = \max[d_1(x_1, y_1), \ldots, d_n(x_n, y_n)]$$

Euclidean n-space can be constructed as the product of n copies of $\mathbb{R}$, each with the Euclidean metric and using the p-norm with $p = 2$.

Consider a product of metric spaces $X_1 \times X_2 \times \ldots \times X_n$. Let $\{x_i^{(j)}\}$ be an infinite sequence in component metric space X_i. Let us call the follow statement "product metric condition #1".

> A sequence $\left(x_1^{(j_1)}, x_2^{(j_2)}, \dots, x_n^{(j_n)}\right)$ converges to $(x_1, x_2, \dots, x_n)$ if and only if $x_i^{(j_i)}$ converges to x_i for $i = 1,2, \dots, n$.

Product metric condition #1 holds true for all the p-norms.

If the metric that we define over a product of metric spaces satisfies product metric condition #1, then the following theorem gives us a way to determine the open sets of the product space.

Theorem 83. If product metric condition #1 holds true for metric $\boldsymbol{d}$ on the product of metric spaces $\boldsymbol{X = X_1 \times X_2 \times \dots \times X_n}$, then the open sets in $\boldsymbol{(X, d)}$ are the unions of product sets of the form $\boldsymbol{Y_1 \times Y_2 \times \dots \times Y_n}$, where $\boldsymbol{Y_i}$ is an open subset of $\boldsymbol{X_i, i = 1, 2, \dots, n}$.

Proof: See Theorem 4.1 in the book "Introduction to Topology" [71]. ■

Given metric spaces $(X_1, d_1), (X_2, d_2), \dots, (X_n, d_n)$ and product metric space $(X_1 \times X_2 \times \dots \times X_n, d)$, we refer to the following property as "product metric condition #2"

$$d_i(x_i, y_i) \leq d(x, y), \qquad x, y \in X, \qquad i = 1,2, \dots, n$$

Product metric condition #2 holds true for all the p-norms.

Theorem 84. If $\boldsymbol{(X_1, d_1), (X_2, d_2), \dots, (X_n, d_n)}$ are complete metric spaces, and the metric $\boldsymbol{d}$ on $\boldsymbol{X_1 \times X_2 \times \dots \times X_n}$ satisfies product metric conditions #1 and #2, then $\boldsymbol{(X, d)}$ is a complete metric space.

Proof: See Theorem 4.2 in the book "Introduction to Topology" [71]. ■

3.4.6 Compactness

Compactness is a property that generalizes the notions of closed and bounded sets from Euclidean space.

From the Wikipedia article "Compact space" [72]:

> In mathematics, specifically general topology, compactness is a property that seeks to generalize the notion of a closed and bounded subset of Euclidean space. The idea is that a compact space has no "punctures" or "missing endpoints", i.e., it includes all limiting values of points. For example, the open interval $(0,1)$ would not be compact because it excludes the limiting values of 0 and 1, whereas the closed interval $[0,1]$ would be compact. Similarly, the space of rational numbers $\mathbb{Q}$ is not compact, because it has infinitely many "punctures" corresponding to the irrational numbers, and the space of real numbers $\mathbb{R}$ is not compact either, because it excludes the two limiting values ∞ and $-\infty$. However, the extended real number line would be compact, since it contains both infinities.

The term "punctures" in the above description is a bit misleading. The idea is to avoid subsets of a metric space that have sequences that converge but not within the subset. For example, the set S consisting of the unit disk (centered at the origin) with the point $\left(\frac{1}{2}, \frac{1}{2}\right)$ removed is not compact since there are many sequences within S that converge to $\left(\frac{1}{2}, \frac{1}{2}\right)$ which, as defined, is not in S. For

example, the points of the sequence $\left\{\frac{1}{2}-\frac{1}{2^n},\frac{1}{2}-\frac{1}{2^n}\right\}, n=0,1,2,3,\ldots$ are within S but the sequence converges to $\left(\frac{1}{2},\frac{1}{2}\right)\notin S$, see Figure 53.

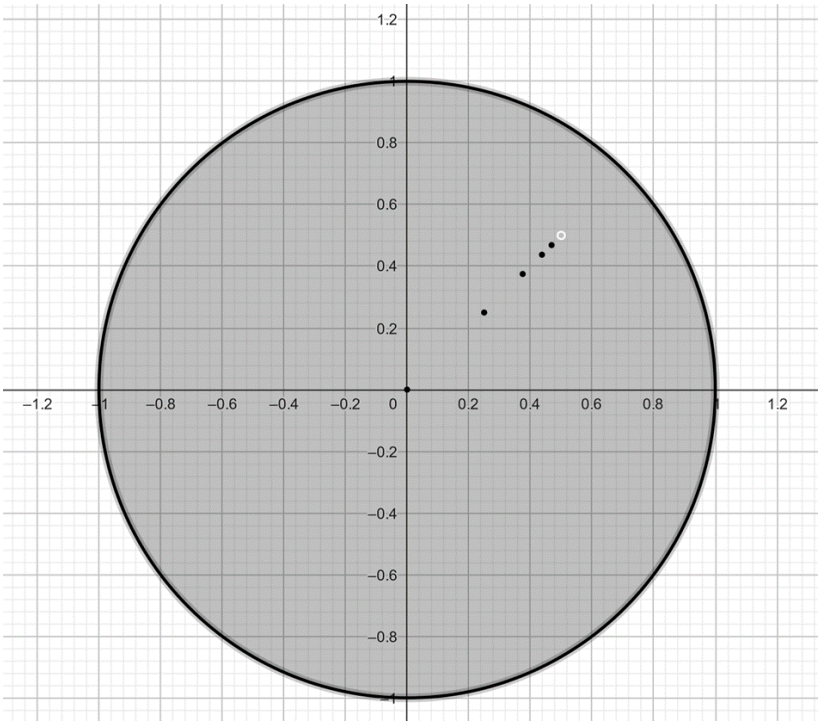


Figure 53. Not a compact set

On the left of Figure 54, we have a disk with the interior of two smaller disks removed. The remaining set (gray area) includes all its boundaries, and is a compact set. The point is that a set can have holes and still be a compact set in some cases.

On the right of Figure 54, the interior of disk A is removed from the larger disk, and the interior and boundary of disk B is removed. The remaining set (gray area) no longer contains all its boundaries and is not a compact set.

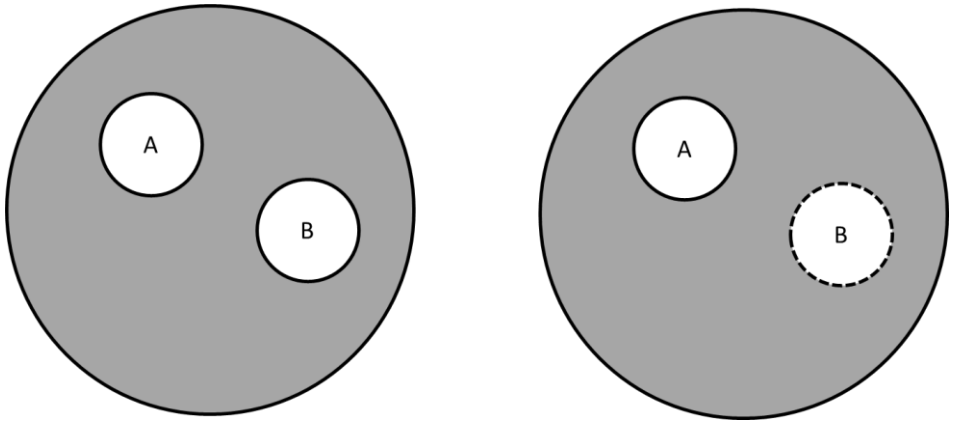


Figure 54. Disks with puncture holes

The formal definition of compactness is a bit technical (details provided below). So, it is recommended that the reader view the YouTube video "The Concept So Much of Modern Math is Built On: Compactness" [73] before proceeding.

A metric space (X, d) is defined to be **compact** if every open cover of X has a finite subcover. In terms of set notation, if

$$X \subseteq \bigcup_{e \in E} A_e$$

where the A_e are open sets (possibly an uncountable collection), then there exist a finite number of the A_e (say $A_1, A_2, \dots, A_n$) such that

$$X \subseteq \bigcup_{i=1}^{n} A_i$$

Warning: this is not saying that X is compact if it has a finite cover of open sets (which is always true).

The definition of compact, as stated above, can be hard to apply. For example, it is far from clear which of the configurations in Figure 54 are compact or not using this definition. Fortunately, there are several other equivalent definitions of compact which may be easier to apply in some situations. In particular, we have the following theorem.

Theorem 85. The following properties are equivalent for a metric space (X, d):

- Every open cover of X has a finite subcover.
- Every sequence in X has a convergent subsequence.
- X is totally bounded and complete. [A metric space X is **totally bounded** if for each $\epsilon > 0$, there exists a finite number of open balls of radius ϵ that cover X.]

Proof: See Theorem 5.1 in the book "Introduction to Topology" [71]. ■

A subset Y of of a metric space (X, d) is said to be a compact subset of (X, d) if every open cover of Y in X has a finite subcover.

The definition for "bounded" is straightforward for metric spaces, i.e.,

> A metric space X is **bounded** if there exists $M > 0$ such that $d(x, y) < M$ for every $x, y \in X$.

As one might expect from the name selection, totally bounded is a more stringent condition than bounded, and in fact, that is true (as demonstrated in the following theorem).

Theorem 86. A totally bounded metric space (X, d) is also bounded.

Proof: Assume X it totally bounded and choose $\epsilon = 1$. By definition, there exists open balls $B(x_i; 1), i = 1,2, \dots, n$ such that

$$X \subseteq \bigcup_{i=1}^{n} B(x_i; 1)$$

Choose $M = 2 + \max\{d(x_i, x_j) : 1 \leq i, j \leq n\}$ which is the maximum distance between the centers of two balls (from the set defined above).

Take any $x, y \in X$.

If x and y are in the same open ball, $d(x,y) < 2 < M$.

If x and y are in different open balls, say $x \in B(x_i;1)$ and $y \in B(x_j;1)$, then

$$\begin{aligned} d(x,y) &\le d(x,x_i) + d(x_i,y) \\ &\le d(x,x_i) + d(x_i,x_j) + d(x_j,y) \\ &< 2 + \max\{d(x_i,x_j): 1 \le i,j \le n\} = M \end{aligned}$$

Either way, $d(x,y) < M$ for any $x, y \in X$. ■

The converse of the above theorem is not true. For example, consider the set of natural numbers $\mathbb{N} = \{1,2,3,\dots\}$ with the discrete metric. Recall that in the discrete metric, the distance between any two distinct points is 1, and the distance between a point and itself is 0.

The set $\mathbb{N}$ is bounded, because for any two elements in the set, the distance is always 1, no matter how far apart you move within the set.

On the other hand, the set is not totally bounded. To see this, imagine we try to cover the entire set with balls of radius $\epsilon < 1$. Since any ball centered at a specific number only contains that single number (since the distance between any two points is at least 1), we would need infinitely many balls of radius ϵ to cover all the natural numbers. This violates the definition of total boundedness, where a finite number of balls should suffice for any chosen value of ϵ.

As another example, consider the set of points Y in infinite dimensional Euclidean space that have a 1 in the i^{th} position and 0 in all other positions, using the following metric:

$$d_\infty\big((x_1,x_2,\dots),(y_1,y_2,\dots)\big) = \max[d_1(x_1,y_1), d_2(x_2,y_2),\dots]$$

The set Y is bounded since the distance between any two of its points is 1. For example, we can take $M = 2$. However, if we choose a radius $\epsilon < 1$, then it takes an infinite number of open balls to cover all the points in Y, and thus, Y is not totally bounded.

In the case of a subset of $\mathbb{R}^n$ under the Euclidean metric, total boundedness is not needed.

Theorem 87. (Heine-Borel theorem) For a subset E of $\mathbb{R}^n$ under the Euclidean metric, the following conditions are equivalent:

- E is compact, i.e., every open cover of E has a finite subcover.
- Every sequence in E has a convergent subsequence.
- E is closed and bounded.

Proof: See the Wikipedia article entitled "Heine–Borel theorem" [74]. ■

3.4.7 Continuous Functions

There are several equivalent definitions of "continuity" for metric spaces. We take the following definition as a basis and then prove two other definitions are equivalent to our base definition.

> Given two metric spaces (X,d) and $Y,d')$, a function $f: X \to Y$ is **continuous at a point** $x \in X$ if whenever $x_n \to x$ for a sequence $\{x_n\}$ in X, then $f(x_n) \to f(x)$.

Implicit in the above definition is that $f(x)$ is defined.

If a function $f: X \to Y$ is continuous at every $x \in X$, then the function f is said to be continuous.

In the following theorem, we prove the first equivalent definition of continuous.

Theorem 88. Given two metric spaces (X, d) and $Y, d')$, and a function $f: X \to Y$ such that $f(x)$ is defined at point $x \in X$. f is continuous at $x \in X$ if and only if for each $\epsilon > 0$, there exist $\delta > 0$ such that whenever $z \in X$ satisfies $d(x, z) < \delta$, then $d'\big(f(x), f(z)\big) < \epsilon$.

Proof: Recall that the statement "A if and only if B" is equivalent to "not A if and only if not B". We will use the latter version in our proof. Further, let A represent the first part of the theorem (i.e., f is continuous) and B represent the condition in the second part of the theorem).

[not B implies not A] Assume there exists $\epsilon > 0$ such that for every $\delta > 0$, there exists a $z \in X$ such that $d(x, z) < \delta$ but $d'\big(f(x), f(z)\big) \geq \epsilon$. Take $\delta = \frac{1}{k}$ and corresponding point z_k such that $d(x, z_k) < \frac{1}{k}$ and $d'\big(f(x), f(z_k)\big) \geq \epsilon$. We have that $z_k \to x$ but $f(z_k)$ does not converge to $f(x)$. So, f is not continuous at x.

[not A implies not B] Going in the other direction, assume f is not continuous at point x. This implies there exists a sequence $\{x_k\}, k = 1,2,3, \ldots$ such that $x_k \to x$ while $f(x_k)$ does not converge to $f(x)$. $f(x_k)$ not converging to $f(x)$implies

$$\exists \epsilon_0 > 0, \forall N \in \mathbb{N}, \exists n_N \geq N \ni d'\left(f\left(x_{n_N}\right) - f(x)\right) > \epsilon_0$$

Taking ϵ_0 from the above expression, choose $N = 1,2,3, \ldots$ which gives us an associated n_N such that $d'\left(f\left(x_{n_N}\right) - f(x)\right) > \epsilon_0$. Since $x_k \to x$, there exists an N such that $d\left(x_{n_N}, x\right) < \delta_{\epsilon_0}$ (i.e., the δ associated with ϵ_0) but $d'\left(f\left(x_{n_N}\right) - f(x)\right) > \epsilon_0$. So, the condition B in the theorem is violated.

■

[**Author's Remark**: The negation of the definition of a convergent sequence (which we used in the second part of the above proof) is complex. The following is an explanation from ChatGPT which I believe is correct.

"A sequence $\{a_n\}$ is said to converge to a limit L if:

$$\forall \epsilon > 0, \exists N \in \mathbb{N} \text{ such that } n \geq N \Longrightarrow |a_n - L| < \epsilon$$

The negation of this statement would be the assertion that the sequence does not converge to L. Let's carefully negate each part of the definition:

1. The original statement starts with "$\forall \epsilon > 0$" which means "for all $\epsilon > 0$."
2. The negation of this would be "$\exists \epsilon > 0$" which means "there exists some $\epsilon > 0$."
3. Next, the original statement says "$\exists N \in \mathbb{N}$" meaning "there exists an $N \in \mathbb{N}$."
4. The negation would be "$\forall N \in \mathbb{N}$" meaning "for all $N \in \mathbb{N}$."
5. Finally, the original statement has "$\forall n \geq N, |a_n - L| < \epsilon$."
6. The negation of this would be "$\exists n \geq N$ such that $|a_n - L| \geq \epsilon$."

Putting these negations together, the negation of the definition of a convergent sequence is:

$$\exists \epsilon > 0, \forall N \in \mathbb{N}, \exists n \geq N \text{ such that } |a_n - L| \geq \epsilon$$

In other words, there exists some positive distance ϵ such that no matter how far out in the sequence you go (no matter how large $N$is), there will always be some terms of the sequence (for $n \geq N$) that stay at least ϵ away from L. This formalizes the idea that the sequence does not converge to the limit L."

end of author's remark.]

We can restate Theorem 88 in terms of open balls as follows:

> f is continuous at $x \in X$ if and only if for each $\epsilon > 0$, there exist $\delta > 0$ such that whenever $z \in B(x; \delta)$, then $f(z) \in B(f(x); \epsilon)$.

or even more concisely,

> f is continuous at $x \in X$ if and only if for each $\epsilon > 0$, there exist $\delta > 0$ such that $f\big(B(x; \delta)\big) \in B(f(x); \epsilon)$.

The following theorem provides yet another equivalent definition of a continuous function.

Theorem 89. Given two metric spaces (X, d) and $Y, d')$, and a function $f: X \to Y$ such that $f(x)$ is defined at point $x \in X$. f is continuous at $x \in X$ if and only if $f^{-1}(V)$ is an open subset of X for every open subset V of Y.

Proof: Assume f is continuous. Let V be an open subset of Y and take any $x \in f^{-1}(V)$. Since V is open there exists $\epsilon > 0$ such that $B(f(x); \epsilon) \subset V$. By the second version of our rewrite of Theorem 88, we have that there exists $\delta > 0$ such that $f\big(B(x; \delta)\big) \subset B(f(x); \varepsilon) \subset V$ which implies $B(x; \delta) \subset f^{-1}(V)$. Since we choose x arbitrarily, $f^{-1}(V)$ contains an open ball about each of its points, and thus, $f^{-1}(V)$ is open.

Going in the other direction, assume $f^{-1}(V)$ is an open subset of X for every open subset V of Y. Take any $x \in X$, and $\epsilon > 0$. Since $B(f(x); \epsilon)$ is an open set in Y, $f^{-1}\big(B(f(x); \epsilon)\big)$ is open in X by hypothesis. Since $f^{-1}\big(B(f(x); \epsilon)\big)$ is open and contains x, there exists $\delta > 0$ such that $B(x; \delta) \subset f^{-1}\big(B(f(x); \epsilon)\big)$. Applying f to both sides of the previous expression, we get $f\big(B(x; \delta)\big) \subset B(f(x); \epsilon)$ which is the condition in the second version of our rewrite of Theorem 88. Thus, f is continuous. ■

Recall that we previously used the term homeomorphism. We are now in a position to formally define a homeomorphism between two metrics spaces X and Y.

A function $f: X \to Y$ between two metric spaces is a **homeomorphism** if it has the following properties:

- f is a bijection (one-to-one and onto). ["onto" means that $\{f(x): x \in X\} = Y$ as opposed to being a proper subset of Y. "one-to-one" means that if $f(x) = f(y)$ then x must equal y.]
- f is continuous.
- The inverse function f^{-1} is continuous.

. . .

There is a condition that is stronger than continuity known as **uniform continuity**. Given two metric spaces (X, d) and $Y, d')$, a function $f: X \to Y$ is uniformly continuous if for each $\varepsilon > 0$, there exists $\delta > 0$ such that whenever $x, z \in X$ satisfy $d(x, z) < \delta$, then $d'(f(x), f(z)) < \varepsilon$. The definition of uniform continuity is the same as that for continuity except that one $\delta > 0$ must hold for all $x \in X$. Clearly, every uniformly continuous function is continuous.

The following theorem exhibits one of many advantages in working with metric spaces that are compact.

Theorem 90. (Heine–Cantor theorem) If X and Y are metric spaces, and X is compact, then every continuous function f from X to Y is uniformly continuous.

Proof: See the Wikipedia article "Heine–Cantor theorem" [75].

3.5 Topological Spaces

It's Deja Vu All Over Again – Yogi Berra

3.5.1 Basic Definitions

A **topology** on a set X is defined by collectionτ of subsets of X, called open sets which satisfy the following axioms:

- The empty set and X belong to τ.
- Any arbitrary (finite or infinite) union of members of τ belongs to τ.
- The intersection of any finite number of members of τ belongs to τ

The set X and the associated topology τ are denoted by (X, τ) and referred to as a **topological space**.

A subset $C \subseteq X$ is said to be **closed** in (X, τ) if its complement $X \backslash C$ is an open set. This implies that X and $\emptyset$ are both open and closed.

A metric space (X, d) induces a topology X called the metric topology. The open sets in a metric topology are X itself, the empty and the open sets as generated by the metric space. Theorem 64 gives us arbitrary union of open sets property, and Theorem 65 gives us the finite union of open sets property.

Not all topological spaces are metrizable (i.e., can be turned into a metric space by adding a metric). The **long line** is an example of a topological space that is not metrizable. It consists of two long rays pointed in opposite directions, where a long ray is an uncountable collection of the interval $[0,1)$ lined up one after the other. More formally, a long ray is defined as the Cartesian product of the first uncountable ordinal $\omega_1$with the half-open interval $[0,1)$, equipped with the order topology that arises from the lexicographical order on $\omega_1 \times [0,1)$. The explanation of why the long line is not metrizable involves concepts beyond the scope of this book, see the Wikipedia article “Long line (topology” [76] for the details.

[**Author’s Remarks**: Given that the theory of topological spaces covers metric spaces, I could have skipped metric spaces and gone directly to topological spaces, and introduced metric spaces as a special case of topological spaces. I did not do this since for one’s first exposure to this topic, metric spaces are already a bit abstract. So, I thought it best to start with something more concrete (e.g., having the concept of distance) and then follow with the more abstract topic of topological spaces.]

A set X may have several possible topologies. Among the possible topologies for a set X are two trivial ones, i.e.,

- the **discrete topology**, consisting of all subsets of X (known as the power set of X)
- the **indiscrete topology**, consisting of only X and the empty set $\emptyset$.

The indiscrete topology for X is not metrizable if X has more than one point. To see this note that the complement of a one-point set $\{p\}$ in a metric space X consisting of more than one point is always open by Theorem 74. Thus, $X \backslash \{p\}$ is open and it does not equal X or $\emptyset$, yielding a contradiction to the assumption that an indiscrete topology on a set of more than one point is metrizable.

3.5.2 Open and Closed Sets

A subset Y of a topological space (X, τ) is a **neighborhood** of a point $x \in X$ if there is an open set U such that $x \in U$ and $U \subseteq Y$. So, Y does not need to be an open set for it to be a neighborhood of a point x; it just needs to contain an open set that, in turn, contains x.

A point $x \in X$ is an **interior point** of Y if Y is a neighborhood of x. The set of interior points of Y is called the interior of Y and is written $int(Y)$. Clearly, $int(Y) \subseteq Y$.

Theorem 91. A subset Y of a topological space (X, τ) is open if and only if $Y = int(Y)$.

Proof: In any event, it is always true that $int(Y) \subseteq Y$.

Assume Y is open, then for every $y \in Y$, Y itself is an open set containing y. Thus, every $y \in Y$ is an interior point of Y, i.e., $Y \subseteq int(Y)$ and thus, $Y = int(Y)$.

Assume $Y = int(Y)$, then for each $y \in Y$, there exist an open neighborhood U_y of y such that $U_y \subseteq Y$. The union of all such U_y is open (property of a topological space) and $\bigcup_{y \in Y} U_y = Y$. Thus, Y is open. ■

Theorem 92. If Y is a subset of a topological space (X, τ), then $int(int(Y) = int(Y)$, i.e., the interior of a subset is open.

Proof: As stated previously, the interior of a subset is contained in the subset itself. For the problem at hand, this means $int(int(Y) \subseteq int(Y)$.

Take any $y \in int(Y)$. By definition, there is an open neighborhood U of y such that $U \subseteq Y$. Since U is an open neighborhood of each of its points, $U \subseteq int(Y)$. Thus, y is an interior point of $int(Y)$, i.e., $y \in int(int(Y))$ which implies $int(Y) \subseteq int(int(Y))$. ■

[**Author's Remark**: As one can see, we are proving theorems analogous to those for metric spaces, but in a more general setting. This pattern will continue throughout most of this section.]

Given a topological space (X, τ), point $x \in X$ is **adherent** to a subset Y of X if every neighborhood of x intersects Y. Each point of Y is then adherent to Y.

The **closure** of Y, written as $\overline{Y}$, is the set of points in X which are adherent to Y. Clearly, $Y \subseteq \overline{Y}$.

Theorem 93. A subset Y of a topological space (X, τ) is closed if and only if $Y = \overline{Y}$.

Proof: In any case, $Y \subseteq \overline{Y}$.

Assume Y is closed. By definition, $X \backslash Y$ is open. Since $X \backslash Y$ is open and $X \backslash Y$ does not intersection Y, no point of $X \backslash Y$ is adherent to Y. Thus, $\overline{Y} \subseteq Y$ which implies $\overline{Y} = Y$.

Going in the other direction, assume $\overline{Y} = Y$ and take $x \in X \backslash Y$. Since $x \notin \overline{Y = Y}$, there exists an open neighborhood U_x of x such $U_x \cap Y = \emptyset$. Taking the union of the U_x, we have $\cup\, U_x = X \backslash Y$. Since the union of open sets is open, $X \backslash Y$ is open and thus, Y is closed. ■

Theorem 94. If Y is a subset of a topological space (X, τ), then $\overline{Y}$ is closed, i.e., $\overline{Y} = \overline{\overline{Y}}$.

Proof: If $x \in X \backslash \overline{Y}$, then then exists an open neighborhood of U_x of x that does not intersect Y. Since U_x is a neighborhood of each of its points, no point of U_x is adherent to $\overline{Y}$, and so, $U_x \subset X \backslash \overline{Y}$. Taking the union of all the U_x, we have

$$\bigcup_{x \in X \backslash \overline{Y}} U_x = X \backslash \overline{Y}$$

So, $X \backslash \overline{Y}$ is open (being the union of open sets) and by definition, its complement $\overline{Y}$ is closed. ■

Even without a metric, we can still define the **convergence of a sequence**. Given a topological space (X, τ), a sequence of points $\{x_i\}$ converges to $x \in X$ if, for every open neighborhood U of x, there is an integer N such that $x_i \in U$ for every $i > N$.

Theorem 95. Given a subset Y of a topological space (X, τ). If a sequence $\{y_i\}$ in Y converges to y, then $y \in \overline{Y}$.

Proof: By Theorem 94, $X \backslash \overline{Y}$ is open. Since $\{y_i\}$ is external to $X \backslash \overline{Y}$, the sequence cannot converge to any point of $X \backslash \overline{Y}$. ■

. . .

The converse of Theorem 95, which is not true, can be stated as follows:

> If y $\in \overline{Y}$, then there exists a sequence $\{y_i\}$ in Y that converges to y.

We will demonstrate in the following example. However, the converse is true for metric spaces, as was shown in Theorem 69.

Let X be a set with topology τ consisting of the collection of subsets U of X such that $X \backslash U$ is at most countable, together with the empty set $\emptyset$. This is known as the **cocountable topology** [77]. For example, let $X = \mathbb{R}$ (real numbers) and let $\mathbb{I}$ represent the irrational numbers. The set $\mathbb{I}$ is open under the cocountable topology on $\mathbb{R}$ since $\mathbb{R} \backslash \mathbb{I}$ (the set of rational numbers) is countable. As another example, let $V_0 = \{0\}$, then $U_0 = \mathbb{R} \backslash V_0$ (the real numbers with 0 removed) is an open set since $\mathbb{R} \backslash U_0 = V_0$ is a finite set.

(X, τ) is a topological space:

> We are given that $\emptyset \in \tau$.
>
> $X \in \tau$ since $X \backslash X = \emptyset$ is finite.
>
> Take any collection of elements from τ, i.e., $\{U_a\}, a \in A$ where A is a potentially uncountable indexing set. By De Morgan's laws for sets, we have
>
> $$X \backslash \bigcup_{a \in A} U_a = \bigcap_{a \in A} X \backslash U_a$$
>
> The right side of the above equation is clearly at most countable (being the intersection of sets which are at most countable). Thus, any union of sets from τ is also in τ.
>
> Take any finite collection of elements from τ, i.e., $\{U_b\}, b \in B$ where B is a finite indexing set. By De Morgan's laws for sets, we have

$$X\backslash\bigcap_{b\in B} U_b = \bigcup_{b\in B} X\backslash U_b$$

The right side of the above equation is at most countable (being the union of a finite collection of at most countable sets). Thus, any finite union of sets from τ is also in τ.

The only convergent sequences in (X,τ) are those that eventually become constant.

Let $\{x_n\}$ be a convergent sequence in X, and let $x \in X$ be its limit. Let $U = X\backslash\{x_n : x_n \neq x\}$, i.e., U is all of X except for elements in the sequence $\{x_n\}$ that do not equal x. So, $x \in U$. Since $X\backslash U = \{x_n : x_n \neq x\}$ is at most countable, $U \in \tau$, i.e., U is an open set. Since U is open, $x \in U$ and $x_n \to x$, it must be that $x_n \in U$ for some $N \in \mathbb{N}$ and all $n \geq N$. If $\{x_n : x_n \neq x\} = \emptyset$, then $\{x_n\} \in X\backslash U$ and yet converges to $x \in U$ with U being open, which is a contradiction. Thus, it must be that $x_n = x$ for all $n \geq N$.

On the other hand, let $\{x_n\}$ eventually become constant, i.e., $x_n = x$ for all $n \geq N$ for some $N \in \mathbb{N}$. In this case, any open neighborhood of x will eventually contain x_n for $n \geq M$ for some $M \in \mathbb{N}$ (M could be less than N) which implies that $x_n \to x$.

Finally, we get to a counterexample to the converse of Theorem 95.

The closed sets in the cocountable topology are the countable sets, the empty set, and X itself. In general, the empty set and the entire set are open and closed for all topological spaces.

Let V be any "at most countable" set in X. Consider $U = X\backslash V$. We have that $X\backslash U = X\backslash(X\backslash V) = V$ which implies that U is open and thus, V is closed.

Let U be any uncountable set in X. Consider $V = X\backslash U$. We have that $X\backslash V = X\backslash(X\backslash U) = U$ which implies that V is not open and thus, U cannot be closed.

Let $A = \{y\}$ be any single element set. From the above, we know that A is closed and thus, $B = X\backslash A$ is open. It must be that $\overline{B} = X$ since there is no other closed set that can contain the open set B. Consequently, the closure of B contains point y, but y cannot be the limit of a sequence in the set B. [**Explanation**: Let $y \in \overline{B}$ be the limit of a sequence $\{y_i\}$ in open set B. Assume $y \in B$. By the definition of convergence, for every open neighborhood of y (B in this case), there exists an integer N such that $y_i \in B$ for all $i > N$. However, for the topology at hand, we know that all convergent sequences eventually equal the limit, which implies $y \in B$. Thus, we have a contradiction and y is not the limit of a sequence in B.]

If we apply the cocountable topology to $\mathbb{R}$ why (for example) doesn't the sequence $\left\{\frac{1}{n}\right\}, n = 1,2,3, \ldots$ converge? Consider the set $A = \left\{\frac{1}{n}\right\}, n = 1,2,3, \ldots$ and the open set $B = \mathbb{R}\backslash A$. Set B contains 0 but contains no elements from A. Thus, by definition, $\left\{\frac{1}{n}\right\}, n = 1,2,3, \ldots$ does not converge.

. . .

A point $b \in X$ is a **boundary point** of a subset Y of X if b is adherent to Y and $X\backslash Y$. The boundary of Y, written as ∂Y, is the set of boundary points of Y, i.e.,

$$\partial Y = \overline{Y} \cap \overline{X\backslash Y}$$

Clearly, $\partial Y = \partial(X\backslash Y)$ and ∂Y is closed (being the intersection of two closed sets).

Theorem 96. If Y is a subset of a topological space (X, τ), then $\overline{Y} = \partial Y \cup int(Y)$ where $\partial Y \cap int(Y) = \emptyset$.

Proof: Take any $y \in \partial Y \cup int(Y) \Longrightarrow y \in int(Y)$ or $y \in \partial Y$. If $y \in int(Y)$, the $y \in int(Y) \subseteq Y \subseteq \overline{Y}$. If $y \in \partial Y$, then $y \in \partial Y \subseteq \overline{Y}$. Either way $y \in \overline{Y}$ and thus, $\partial Y \cup int(Y) \subseteq \overline{Y}$.

Take any $y \in \overline{Y}$. Then every neighborhood U of y intersects Y. There are two mutually disjoint cases, i.e., either there is a neighborhood U of y such that $U \subset Y$, or each neighborhood U of y intersects $X \backslash Y$. The former case occurs if and only if $y \in int(Y)$, and the latter case occurs if and only if $y \in \partial Y$. Thus, $y \in \partial Y \cup int(Y)$ which implies $\overline{Y} \subseteq \partial Y \cup int(Y)$. ■

3.5.3 Subspaces

Given a topological space (X, τ) and a subset S of X, the **subspace topology** on S is defined by

$$\tau_S = \{S \cap U : U \in \tau\}$$

In words, a subset of S is open in the subspace topology if and only if it is the intersection of S with an open set in (X, τ). If S is equipped with the subspace topology then it is a topological space in its own right, and is called a subspace of (X, τ). Subsets of topological spaces are assumed to be equipped with the subspace topology unless otherwise stated.

The sets $V \in \tau_S$ are **relatively open** subsets of S, and the sets $S \backslash V$ (for $V \in \tau_s$) are **relatively closed** subsets of S.

The following are examples subspaces of $\mathbb{R}$ with the standard Euclidean metric (taken from the Wikipedia article "Subspace topology" with some editing):

- The subspace topology of the natural numbers, as a subspace of $\mathbb{R}$, is a discrete topology.
- The rational numbers $\mathbb{Q}$ considered as a subspace of $\mathbb{R}$ do not have the discrete topology. For example, $\{0\}$ is not an open set in $\mathbb{Q}$ because there is no open subset of $\mathbb{R}$ whose intersection with $\mathbb{Q}$ can result in only the singleton $\{0\}$. For $a, b \in \mathbb{Q}$, the intervals (a, b) and $[a, b]$ consisting of only rational numbers are respectively open and closed in $\mathbb{Q}$.
- The set $[0,1]$ as a subspace of $\mathbb{R}$ is both open and closed, whereas as a subset of $\mathbb{R}$ it is only closed.
- As a subspace of $\mathbb{R}$, $[0, 1] \cup [2, 3]$ is composed of two disjoint open subsets (which happen to also be closed), and is therefore a disconnected space.
- Let $S = [0, 1)$ be a subspace of the real line $\mathbb{R}$. For example, $\left[0, \frac{1}{2}\right)$ is open in S but not in $\mathbb{R}$, since the intersection between the open set $\left(-\frac{1}{2}, \frac{1}{2}\right)$ in $\mathbb{R}$ and S results in $\left[0, \frac{1}{2}\right)$. Similarly, $\left[\frac{1}{2}, 1\right)$ is closed in S but not in $\mathbb{R}$. Further, $\left[\frac{1}{2}, 1\right)$ is not open in S since there is no open subset of $\mathbb{R}$ that can intersect with $[0, 1)$ and result in $\left[\frac{1}{2}, 1\right)$. S is both open and closed as a subset of itself but not as a subset of $\mathbb{R}$.

Using the definition of an open set in a subspace, we can derive a similar condition for a set to be closed in a subspace.

Theorem 97. Let S be a subspace of a topological space X. A subset A of S is relatively closed in S if and only if A is the intersection of S and a closed subset of X.

Proof: If A is a relatively closed subset in S then (by definition) $S\backslash A$ is relatively open in S which implies that there exists an open subset B of X such that $S\backslash A = B \cap S$. Thus, A is the intersection of S and the closed subset $X\backslash B$ of X.

Going in the other direction, if $A = T \cap S$ where T is a closed subset of X, then $S\backslash A$ is the intersection of S and the open subset $X\backslash T$ of X. Thus, $S\backslash A$ is relatively open in S, and A is relatively closed in S. ■

3.5.4 Continuous Functions

A function $f: X \to Y$ between topological spaces X and Y is **continuous** if $f^{-1}(V)$ is open in X whenever V is an open subset of Y. The idea is to capture the intuition that there are no "jumps" or "separations" in the function.

A function $f: X \to Y$ is continuous at a point $x \in X$ if for every open set V in Y such that $f(x) \in V$, there exists an open set U in X such that $x \in U$ and $f(U) \subseteq V$.

Theorem 98. Let X and Y be topological spaces. A function $f: X \to Y$ is continuous if and only if f is continuous at each point of X.

Proof: First, assume that f is continuous at each point of X. Take **any** open set V in Y, and take **any** $x \in f^{-1}(V)$. By the definition of continuity at a point, there exists an open set U_x that contains x such that $f(U_x) \subseteq V$ which implies $U_x \subseteq f^{-1}(V)$. Take the union of all the U_x sets, i.e.,

$$U = \bigcup_{x \in f^{-1}(V)} U_x$$

Being the union of open sets, U is open. In addition, $f(U) \subseteq V$ which implies $U \subseteq f^{-1}(V)$. Since U contains every point of $f^{-1}(V)$, we have $f^{-1}(V) \subseteq U$. Thus, $f^{-1}(V) = U$. Finally, $f^{-1}(V)$ being open implies f is continuous.

Conversely, assume f is continuous. Let $x \in X$ and let V be an open set containing $f(x)$, i.e., $f(x) \in V$. The set $U = f^{-1}(V)$ is open since V is open and f is continuous. Further, $x \in U$ and $f(U) \subseteq V$ and thus, f is continuous at x (which was chosen arbitrarily). ■

Given two functions $f: X \to Y$ and $g: Y \to Z$, where X, Y, Z are topological spaces, a third function (known of the **composition of functions**) can be defined. The composition of g with f at x is written $g \circ f(x)$ or $g\big(f(x)\big)$. If $X \neq Z$, then the $f \circ g$ is not defined.

For example, take $X = Y = Z = \mathbb{R}$, $f(x) = x^2$ and $g(x) = e^x$. We have

$$f \circ g(x) = f\big(g(x)\big) = f(e^x) = (e^x)^2 = e^{2x}$$

$$g\big(f(x)\big) = g(x^2) = e^{x^2}$$

The inverse function of a composition (assuming both functions are invertible) has the property that $(f \circ g)^{-1} = g^{-1} \circ f^{-1}$ (for a proof of this fact, see Section 4.3, p. 362 of Rodgers [83]).

Theorem 99, If $f: X \to Y$ and $g: Y \to Z$ are continuous and X, Y, Z are topological spaces, then $g \circ f$ is continuous.

Proof: Take any open set V in Z. Since g is continuous, $g^{-1}(V)$ is open in Y. Since f is continuous, $f^{-1}\left(g^{-1}(V)\right) = (g \circ f)^{-1}(V)$ is open in X. Thus, $g \circ f$ is continuous. ■

This is much easier than the ε/δ type of proof used to prove the result for metric spaces. Yet this result and the associated proof holds for metric spaces.

. . .

A **homeomorphism** between topological spaces is a bijection (one-to-one and onto function) that is continuous and whose inverse is also continuous. From the standpoint of topology, homeomorphic spaces are essentially identical.

A characteristic of a topological space qualifies as a topological property if it remains unchanged under homeomorphisms. For instance, discreteness exemplifies a topological property: if two spaces, X and Y, are homeomorphic, X is discrete if and only if Y is discrete. Metrizability also constitutes a topological property, though attributes tied to a particular metric usually fall outside this category. Broadly speaking, any property describable in relation to the open or closed subsets of a space qualifies as a topological property.

For example, take the open intervals (a, b) and (c, d) in $\mathbb{R}$. Each interval is a subspace of $\mathbb{R}$. Consider the function $f(x) = \left(\frac{d-c}{b-a}\right)(x - a) + c, a < x < b$. f is continuous (it is just a line segment), $f(a) = c, f(b) = d$ and for $a < x < b, c < f(x) < d$. So, topological spaces (a, b) and (c, d) are homeomorphic.

3.5.5 Base for a Topology

A **base** (or basis) for the topology τ of a topological space (X, τ) is a collection $\mathcal{B}$ taken from τ such that every open set of the topology is equal to the union of some elements of $\mathcal{B}$. For example, the set of all open intervals in the real number line $\mathbb{R}$ is a basis for the Euclidean topology on $\mathbb{R}$ because every open interval is an open set, and every open subset of $\mathbb{R}$ can be written as a union of open intervals.

Bases are ubiquitous throughout topology. The sets in a base for a topology, which are called **basic open sets**, are often easier to describe and use than arbitrary open sets. For example, the open balls in a metric space (which do, in fact, form a base) are easy to describe and use versus the collection of all possible open sets. Many important topological definitions such as continuity and convergence can be checked using only basic open sets instead of arbitrary open sets.

In general, a topological space (X, τ) can have several different bases. The whole topology τ is always a base for itself (but also the most verbose base). Further, two different bases need not have any basic open set in common.

The following theorem gives a condition for a collection of open subsets of a topological space to be a base.

Theorem 100. A collection $\mathcal{B}$ of open subsets of a topological space X is a base for a topology of X if and only if for each $x \in X$ and each neighborhood U of x, there exists $V \in \mathcal{B}$ such that $x \in V$ and $V \subseteq U$.

Proof: If $\mathcal{B}$ is a base of X, then each open neighborhood U of x is a union of sets from $\mathcal{B}$, and so, there must be some $V \in \mathcal{B}$ such that $x \in V \subseteq U$.

Conversely, assume the condition holds true. Take any open subset U of X. For each $x \in U$, there exists $V_x \in \mathcal{B}$ such that $x \in V_x \subseteq U$. Clearly,

$$U = \bigcup_{x \in U} V_x$$

So, any open subset of X is a union of sets from $\mathcal{B}$ and therefore, $\mathcal{B}$ is a base for the topology. ■

As noted, it is sometimes easier to define a topology on a set X by specifying a base for the topology. However, not every collection of subsets of X are a base for a topology. It is necessary to check that the proposed base satisfies the conditions of the following theorem.

Theorem 101. A collection of subsets $\mathcal{B}$ taken from a set X is a base for a topology of X if and only if has the following properties:

(1) Each $x \in X$ is an element of at least one set in $\mathcal{B}$.

(2) If $U, V \in \mathcal{B}$ and $x \in U \cap V$, then there exists $W \in \mathcal{B}$ such that $x \in W$ and $W \subseteq U \cap V$.

Proof: Assume that $\mathcal{B}$ is a base for a topology of set X. Since X itself is open, it must be the union of sets from $\mathcal{B}$ and thus, each $x \in X$ is an element of at least one member of $\mathcal{B}$. This gives us property (1). Property (2) holds since $U \cap V$ is open and is thus the union of members from $\mathcal{B}$, at least one of which must contain x.

Going in the other direction, assume $\mathcal{B}$ is a collection of subsets of X for which properties (1) and (2) hold. Let τ be the collection of all subsets of X that are unions of sets from $\mathcal{B}$, including the empty set. Clearly, $\mathcal{B}$ is a base for X. However, we still need to show that $\mathcal{B}$ is a valid topology.

By property (1), X is the union of all members of $\mathcal{B}$. So, X is in τ.

Clearly, the union of sets in τ is also in τ (given how we defined τ). So, it only remains for us to show that τ is closed under finite intersections. To that end, take $U, V \in \tau$ and take any $x \in U \cap V$. By the definition of τ, U and V are each the union of sets in $\mathcal{B}$, and so, there exists $U_x, V_x \in \mathcal{B}$ such that $x \in U_x \subseteq U$ and $x \in V_x \subseteq V$. Thus, $x \in U_x \cap V_x$. By property (2), there exists $W_x \in \mathcal{B}$ such that $x \in W_x \subseteq U_x \cap V_x$ which implies $W_x \subseteq U \cap V$. Clearly,

$$U \cap V = \bigcup_{x \in U \cap V} W_x$$

Thus, $U \cap V \in \tau$.

This easily extends to a finite collection of sets in τ by intersection two at a time. ■

A **second-countable space**, also known as a completely separable space, is a topological space whose topology has a countable base. More precisely, a topological space (X, τ) is second-

countable if there exists some countable collection $\mathcal{U} = \{U_i\}, i = 1,2, \ldots$ of open subsets of X such that any open subset of X can be written as a union of elements from a subfamily of $\mathcal{U}$. A second-countable space is said to satisfy the second axiom of countability.

Many "well-behaved" spaces in mathematics are second-countable. For example, Euclidean space $\mathbb{R}^n$ with its usual topology of open balls is second-countable. Although the base of open balls is uncountable, one can restrict the collection to all open balls with rational radii and whose centers have rational coordinates. This restricted set is countable and still forms a basis of $\mathbb{R}^n$.

We have the following definitions for topological space each of which is very similar to the analogous definition for metric spaces.

- An **open cover** of a topological space X is a collection of open subsets of X whose union equals X. A subcover of an open cover $\mathcal{C}$ of X is a subset of $\mathcal{C}$ whose union is X.
- A subset Y of a topological space X is **dense** in X if $\overline{Y} = X$. A topological space X is **separable** if there is a countable subset of X that is dense in X.

Second-countability implies other important topological properties, e.g., every second-countable space is separable and Lindelöf (every open cover has a countable subcover).

3.5.6 Separation

At the end of the previous section, we restricted our focus to topological spaces with a base of a certain type, i.e., topologies with a countable base. The property of second-countability effectively requires that a topological space does not have too many open sets. The properties considered in this section require that a topological space does not have too few open sets. The reason for this requirement is that topologies with too few open sets tend to provide little information since their structure is overly simple. The extreme example is the indiscrete topology, which provides no structural information at all.

Warning: The term “separable” as used in the previous section, and the terms “separation” and “separated” in this section have different meanings. Unfortunately, these meanings are embedded in the mathematical literature.

The following is a short list of the many separation axioms for topological spaces. A more complete list can be found in the Wikipedia article “Separation axiom” [79]. The separation axioms are denoted with the letter "T" after the German Trennungsaxiom ("separation axiom"), and increasing numerical subscripts denote stronger and stronger properties.

- Two points of a topological space X are **topologically indistinguishable** if they have exactly the same neighborhoods. Two points of X are **topologically distinguishable** if they are not topologically indistinguishable. For example, $x \in X$ could have a neighborhood that does not include $y \in X$, but every neighborhood of y could include x. In this case, x and y would be topologically distinguishable.
- In a topological space X, points x and y are **separated** if each of them has a neighborhood that is not a neighborhood of the other, i.e., neither belongs to the closure of the other.
    - Two subsets A and B of X are separated if each is disjoint from the other's closure, though the closures themselves do not have to be disjoint. The neighborhoods don’t need to be disjoint.

  - Suppose A and B are subspaces of topological space X. A and B said to be separated if each is disjoint from the closure of the other.

- A topological space X is T_0, or Kolmogorov, if any two distinct points in X are topologically distinguishable.
- A topological space X is T_1, or accessible or Fréchet, if any two distinct points in X are separated.
- A topological space X is **Hausdorff**, or T_2 or separated, if any two distinct points in X are separated by disjoint neighborhoods. Every Hausdorff space is also T_1. The Hausdorff property is designed to make limits of sequences unique.
- A topological space X is **regular** if any point x and any closed set F in X (but not containing x) are separated by neighborhoods.
- A topological space X is regular Hausdorff, or T_3, if it is T_0 and regular.
- A topological space X is **normal** if any two disjoint closed subsets of X are separated by neighborhoods.
- A topological space X is normal Hausdorff, orT_4, if it is T_1 and normal.

Figure 55 depicts examples of the Hausdorff, regular and normal properties in $\mathbb{R}^2$. E and F are closed sets, and x and y are points. The containing shapes (light gray) are neighborhoods.

To summarize the relationships, we have that $T_4 \Longrightarrow T_3 \Longrightarrow T_2 \Longrightarrow T_1$.

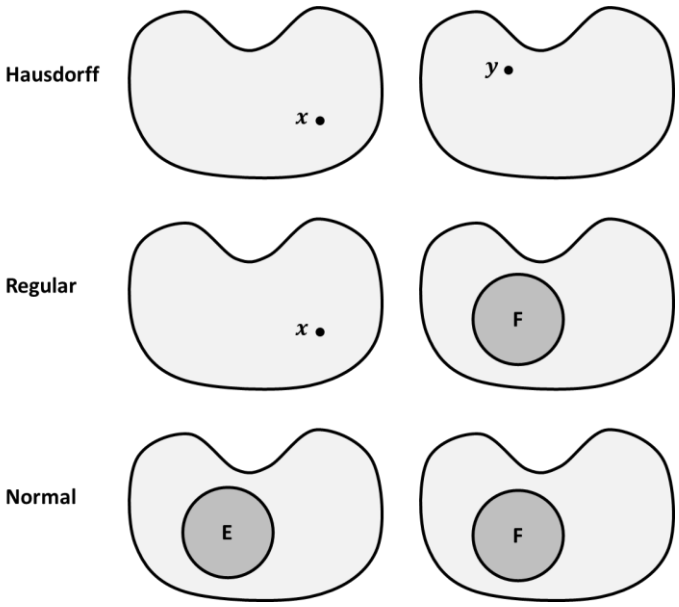


Figure 55. Example of Hausdorff, regular and normal properties

The following theorem gives us another way to characterize $T1$ spaces.

Theorem 102. A topological space X is a $T1$ space if and only if each of its points are closed sets.

Proof: Assume that X is a $T1$ space. Taken any $x \in X$. For every $y \neq x$, there exists an open set U_y such that $y \in U_y$ and $x \notin U_y$. Let $U = \bigcup_{y \neq x} U_y$ which is open, being the union of open sets. So, $X \backslash \{x\} = U$ is open and thus, $\{x\}$ is closed.

Conversely, if $\{x\}$ is a closed subset of X, then $U = X \backslash \{x\}$ is an open set containing each $y \in X, y \neq x$. ■

The following theorem tells us where metric spaces fit into this scheme.

Theorem 103. Every metric space is a T_4 space.

Proof: Assume that (X, d) is a given metric space. Since each point $\{x\}$ can been represented as the intersection of all closed sets $\overline{B(x; r)}, r > 0$, each point in X is a closed set, and thus, X is a T_1 space by Theorem 102.

Concerning normality condition, take any two disjoint closed sets E and F in X. For each $x \in E$, there exists $r_x > 0$ such that $B(x; r_x) \cap F = \emptyset$, and for each $y \in F$, there exists r_y such that $B(y; r_y) \cap E = \emptyset$. Next, define the following sets

$$U = \bigcup_{x \in A} B\left(x; \frac{r_x}{2}\right)$$

$$V = \bigcup_{y \in B} B\left(y; \frac{r_y}{2}\right)$$

It follows that $E \subset U$ and $F \subset V$. To complete the proof, we need to show that U and V are disjoint. To that end, assume there exists $z \in U \cap V$. This implies that there exist $x \in E$ and $y \in F$ such that $d(x, z) < \frac{r_x}{2}$ and $d(y, z) < \frac{r_y}{2}$. Thus,

$$d(x, y) \leq d(x, z) + d(y, z) < \frac{r_x}{2} + \frac{r_y}{2} \leq \max(r_x, r_y)$$

If $r_x \geq r_y$, then $y \in B(x; r_x)$, and if $r_x \leq r_y$, then $x \in B(y, r_y)$. Either way, we have a contraction, and so, the intersection of U and V must be empty. ■

. . .

Subsets A and B of a topological space X are said to be **separated by a continuous function** if there exists a continuous function $f: X \to [0, 1]$ from X into the unit interval $[0,1]$ such that $f(a) = 0, \forall a \in A$ and $f(b) = 1, \forall b \in B$. Any such function is called a **Urysohn function** for A and B. In this case, A and B are necessarily disjoint.

For normal topological spaces, we have the following important theorem concerning subsets of a topological space that are separated by a continuous function.

Theorem 104. (Urysohn's lemma) A topological space X is normal if and only if, for any two non-empty closed disjoint subsets A and B of X, there exists a continuous map $f: X \to [0, 1]$ from X into the unit interval $[0, 1]$ such that $f(a) = 0, \forall a \in A$ and $f(b) = 1, \forall b \in B$, i.e., there exists a Urysohn function for A and B.

Proof: The Wikipedia article "Urysohn's lemma" [80] provides a sketch of the proof. ■

Urysohn's lemma is used in the proof of Urysohn's metrization theorem, which gives criteria for a topological space to be metrizable.

Theorem 105. (Urysohn's metrization theorem). Every second-countable regular Hausdorff space is metrizable.

Proof: See Theorem 34.1 of the topology book by Munkres [81]. ■

3.5.7 Compactness

Recall that the definition of compactness for metric spaces relied solely on open sets and not the metric. We can define compactness for topological spaces in the same way, i.e., a topological space is **compact** if every open cover has a finite subcover.

A subset Y of a topological space X is a compact subset of X if Y is compact in the relative topology it inherits from X. This happens if and only if Y has the following property:

> If $\{U_a\}, a \in A$ is a (possibly uncountable) collection of open subsets of X that covers Y, then there is a finite subcollection of the U_a sets that covers Y.

The definition of compactness for topological spaces is cast only in terms of set theory ideas (unions and set inclusions) and the openness of sets. A homeomorphism, being a one-to-one and onto function, preserves unions and inclusions, and it preserves openness by definition. Thus, a homeomorphism between topological spaces preserves compactness.

The following are a couple of basic theorems concerning compact topological spaces.

Theorem 106. A finite union of compact subsets of a topological space is compact.

Proof: Let $\{Y_i\}, i = 1,2, \dots, n$ be a collection of compact subsets of topological space X. Let

$$Y = \bigcup_{i=1}^{n} Y_i$$

Let $U_a, a \in A$ be a (possibly uncountable) collection of open subsets of X that covers Y, and thus, covers each Y_i. For each $i = 1,2, \dots, n$, there is a finite subcollection of the U_a sets that covers Y_i since the Y_i sets are compact. The union of each of the finite subcovers for $Y_i, i = 1,2, \dots, n$ forms a finite subcover for Y. ■

Theorem 107. A closed subspace Y of a compact topological space X is compact.

Proof: Let $\{U_a\}, a \in A$ be a collection of open subsets of X that covers Y. Since Y is closed, $X \backslash Y$ is open. Thus, $U = \{U_a \cap X \backslash Y\}, a \in A$ is an open cover for X. Since X is compact, there is a subcollection from U (call it V) that covers X. If we take the sets in V (with $X \backslash Y$ removed if it is in V), then we have a finite subcover of Y and thus, Y is compact. ■

The following theorems relate compactness to the Hausdorff separation property.

Theorem 108. Let Y be a compact subset of a Hausdorff space X. For each $x \in X \backslash Y$, there exist disjoint open neighborhoods U of x and V of Y.

Proof: Since X has the Hausdorff property, for each $y \in Y$, there exist disjoint open neighborhoods of $x \in X \backslash Y$ and y (call them U_y and V_y, respectively). The open sets $\{V_y\}, \forall y \in Y$ form a cover of Y, and since Y is compact, there is a finite subcover, say $Y \subset V_{y_1} \cup V_{y_2} \cup \dots \cup V_{y_n} = V$.

Taking the corresponding U_{y_i} for x relative to y_i, we have $U_{y_1} \cup U_{y_2} \cup \ldots \cup U_{y_n} = U$ which is an open neighborhood of x that is disjoint from V. ■

Theorem 109. A compact subset Y of a Hausdorff space X is closed.

Proof: By Theorem 108, for every $x \in X \backslash Y$ there are disjoint open neighborhoods of x and Y. So, $X \backslash Y$ is open, being the union of open sets, and thus, Y is closed. ■

Theorem 110. Every compact Hausdorff space is normal.

Proof: Take any closed subsets S and T from compact Hausdorff space X. We know by Theorem 107 that S and T are compact. By Theorem 108, for each $t \in T$, there exists disjoint neighborhoods U_t of t and V_t of S. The open sets $\{U_t\}, t \in T$ provide a cover for T. Since T is compact, it has a finite subcover, i.e.,

$$T \subset U_{t_1} \cup U_{t_2} \cup \ldots \cup U_{t_n} = U$$

If we let $V = V_{t_1} \cap V_{t_2} \cap \ldots \cap V_{t_n}$, then $S \subset V$. By construction, $V \cap U = \emptyset$. Thus, X is normal. ■

A continuous function maps a compact space to a compact space.

Theorem 111. Let f be a continuous function from a compact topological space X to a topological space Y. Then $f(X)$ is a compact subset of Y.

Proof: If $\{U_a\}, a \in A$ is an open cover of $f(X)$, then $\{f^{-1}(U_a)\}, a \in A$ is an open cover of X. Since X is compact, there exists a finite open cover, i.e.,

$$X = f^{-1}\left(U_{a_1}\right) \cup f^{-1}\left(U_{a_2}\right) \cup \ldots \cup f^{-1}\left(U_{a_n}\right)$$

where $a_i \in A, i = 1,2, \ldots, n$.

So, $f(X) = U_{a_1} \cup U_{a_2} \cup \ldots \cup U_{a_n}$, i.e., there is a finite subcover for $f(X)$. Thus, $f(X)$ is compact. ■

3.5.8 Connectedness

Intuitively, a topological space is connected if it is all in one piece. This leads one to conclude that a topological space is disconnected if it can be written as the union of two non-empty "separated" pieces. To make this precise, we need to determine what "separated" should mean. For example, we correctly think of $\mathbb{R}$ as connected even though it can be written as the union of two disjoint piece, e.g., $(-\infty, 1]$ and $(1, \infty)$. So, "separated" must mean something more than "disjoint." In fact, we need to use the definition of separated sets that was stated earlier, i.e., two subsets of a topological space are separated if each is disjoint from the closure of the other.

A topological space X is **connected** if it cannot be expressed as the union of two non-empty subsets that are both open and closed, and separated from each other. On the other hand, a topological space is **disconnected** if there are closed and open subsets U and V of X such that

$$U \cup V = X$$
$$(U \cap \overline{V}) \cup (V \cap \overline{U}) = \emptyset$$
$$U \neq \emptyset, V \neq \emptyset.$$

For a topological space X the following conditions are equivalent [84]:

- X is connected, i.e., it cannot be divided into two disjoint non-empty open sets.
- The only subsets of X which are both open and closed (aka **clopen** sets) are X and the empty set.
- The only subsets of X with empty boundary are X and the empty set.
- X cannot be written as the union of two non-empty separated sets (sets for which each is disjoint from the other's closure).
- All continuous functions from X to $\{0,1\}$ are constant, where $\{0,1\}$ is the two-point space endowed with the discrete topology.

Some examples:

- $\mathbb{R}\backslash\{0\}$ is disconnected since we can write it as the union of two separated open sets, i.e., $(-\infty, 0) \cup (0, \infty)$.
- The closed interval $[-1,1)$ in the standard subspace topology is connected. It can, however, be written as the union of $[-1,0)$ and [0,1) but the second set is not open in the chosen topology.
- The union of $[-1,0)$ and $(0,1]$ is a disconnected topological subspace of $\mathbb{R}$ since [-1,0) and (0,1] are open in the standard topological space $[-1,0) \cup (0,1]$.
- As a topological subspace of $\mathbb{R}$, $(1,2) \cup \{5\}$ is disconnected.

. . .

Path-connectedness is a stronger notion than connectedness as it requires the structure of a path. A path from a point x to a point y in a topological space X is a continuous function f from the unit interval $[0,1]$ to X with $f(0) = x$ and $f(1) = y$. A path-component of X is an equivalence class of X under the equivalence relation which makes x equivalent to y if there is a path from x to y. The space X is said to be path-connected (or pathwise connected or 0-connected) if there is exactly one path-component. For non-empty spaces, this is equivalent to the statement that there is a path joining any two points in X.

A topological space is said to be **simply connected** if it is path-connected and every path between two points can be continuously transformed into any other such path while preserving the two endpoints in question. Intuitively, this corresponds to a space that has no disjoint parts and no holes that go completely through it, because two paths going around different sides of such a hole cannot be continuously transformed into each other.

3.5.9 Return to Surfaces

The concept of a surface (as discussed in Section 3.3) can be related to the various topological concepts that we have developed in this section.

A **manifold** is a topological space that locally resembles Euclidean space near each point. More precisely, an n-dimensional manifold (or n-manifold) is a topological space with the property that each point is surrounded by a neighborhood that is homeomorphic to an open subset of $\mathbb{R}^n$.

One dimensional manifolds include lines and circles, but not self-intersecting curves. Two-dimensional manifolds are known as surfaces. As we have seen, the plane, sphere, torus, Klein bottle and real projective plane are examples of surfaces.

In many but not all cases, it is assumed that a surface is nonempty, second-countable, and Hausdorff. It is also often assumed that the surfaces under consideration are connected.

4 Complex Analysis

4.1 Overview

Recall from high school algebra that the equation $ax^2 + bx + c = 0$ has solutions given by the quadratic formula

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

The solution gives the x-value for the intersection of a parabola with the x-axis. For example, consider the parabola $y = x^2 - 3x + 2 = (x - 2)(x - 1)$. Since we can factor the function, we don't need the quadratic formula to see that the roots (i.e., values of x for which the equation equals 0) are 1 and 2, see Figure 56.

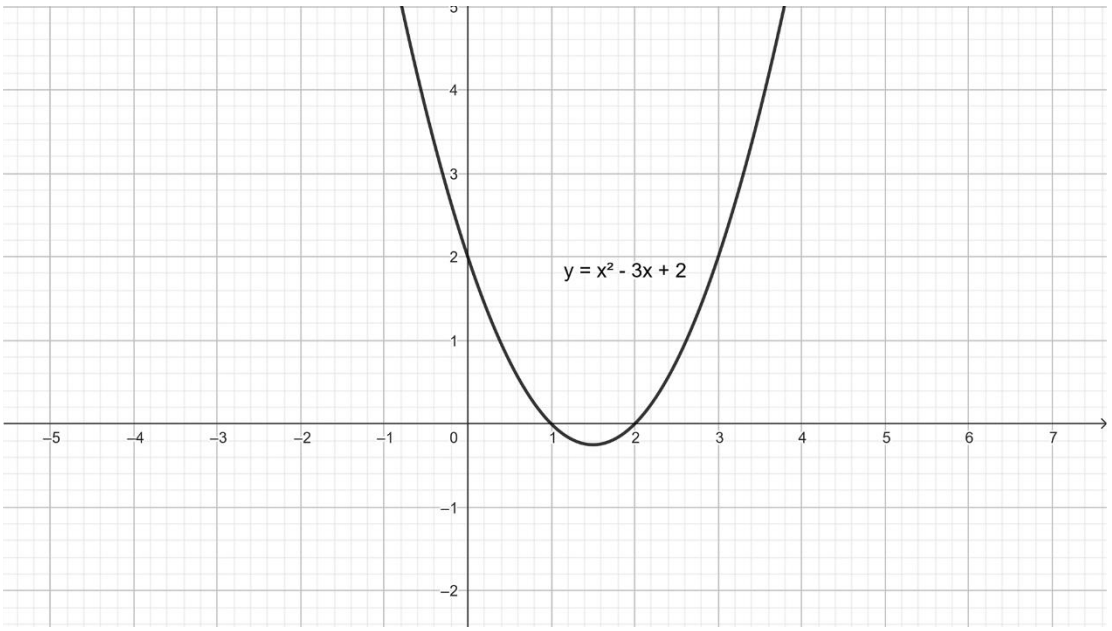


Figure 56. Parabola with two distinct real roots

There are several other cases (i.e., other than having to real-valued roots):

- The quadratic equation has double real roots, e.g., $x^2 - 2x + 1 = (x - 1)^2$ has a double root at $x = 1$.
- The quadratic equation has no real roots, e.g., $x^2 - 3x + 4$ has no real roots since the parabola does not intersect the x-axis, see Figure 57. However, it does have two complex roots, i.e.,

$$\frac{3 \pm \sqrt{-7}}{2} = \frac{3}{2} \pm \frac{\sqrt{-1}\sqrt{7}}{2}$$

In general, for quadratic equations that have no real roots, there will be a $\sqrt{-1}$ term. We use the symbol i to represent $\sqrt{-1}$, noting that $i^2 = -1$. With this simple idea, we have the beginnings of the topic known as complex analysis.

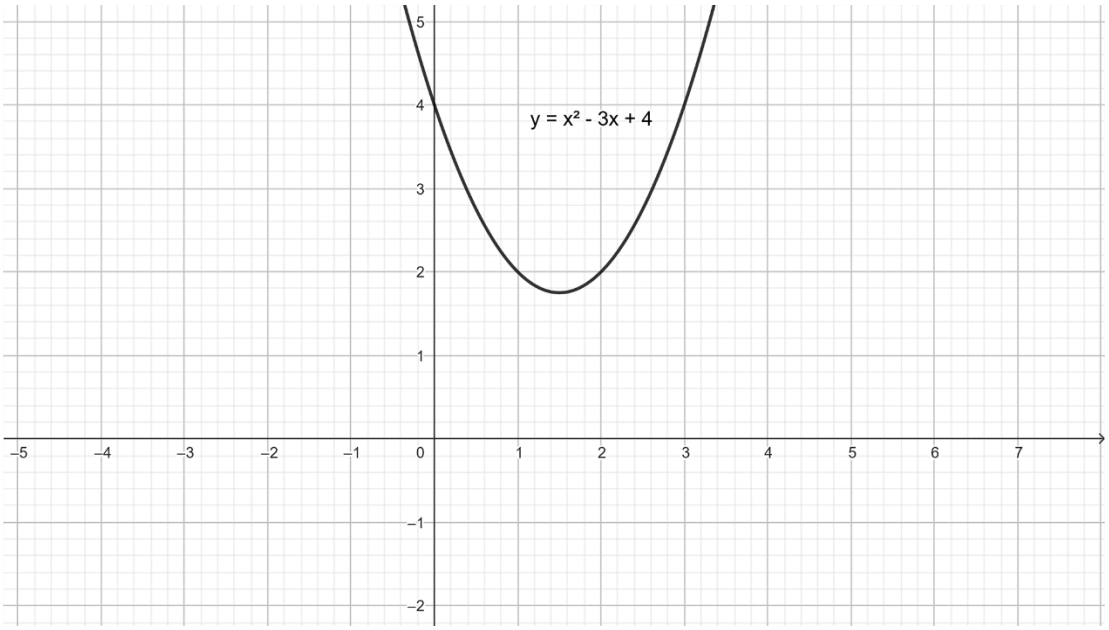


Figure 57. Parabola with no real roots

In studying the roots of quadratic equations, we have discovered a new type of number known as a complex number. Complex numbers have the general form $a + bi$ where $a, b \in \mathbb{R}$ and $i = \sqrt{-1}$ (or $i^2 = -1$). If we let $z = a + bi$ then the **real part** of z, denoted by $Re(z)$, is a and the **imaginary part** of z, denoted $Im(z)$, is b. The set of all complex number is denoted $\mathbb{C}$.

In the following subsections, we will study the properties of complex numbers, functions with arguments from the complex numbers (known as complex functions), derivatives of complex functions and the integration of complex functions.

4.2 Basic Properties of Complex Numbers

The set of all real numbers can be represented by a single axis. To graphically represent the complex numbers, we need two axes (one axis for the real part, referred to as the real axis, and one axis for the imaginary part, referred to as the imaginary axis). Figure 58 shows the complex plane along with three example points. The points are represented two ways, i.e., as a complex number in the form $a + bi$ and as an ordered pair (a, b) where a represents the real part and b represents the imaginary part.

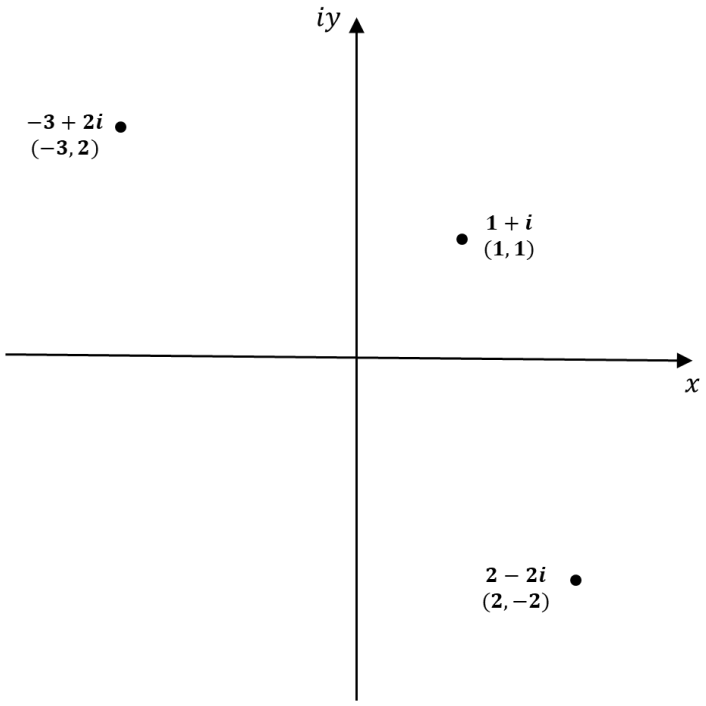
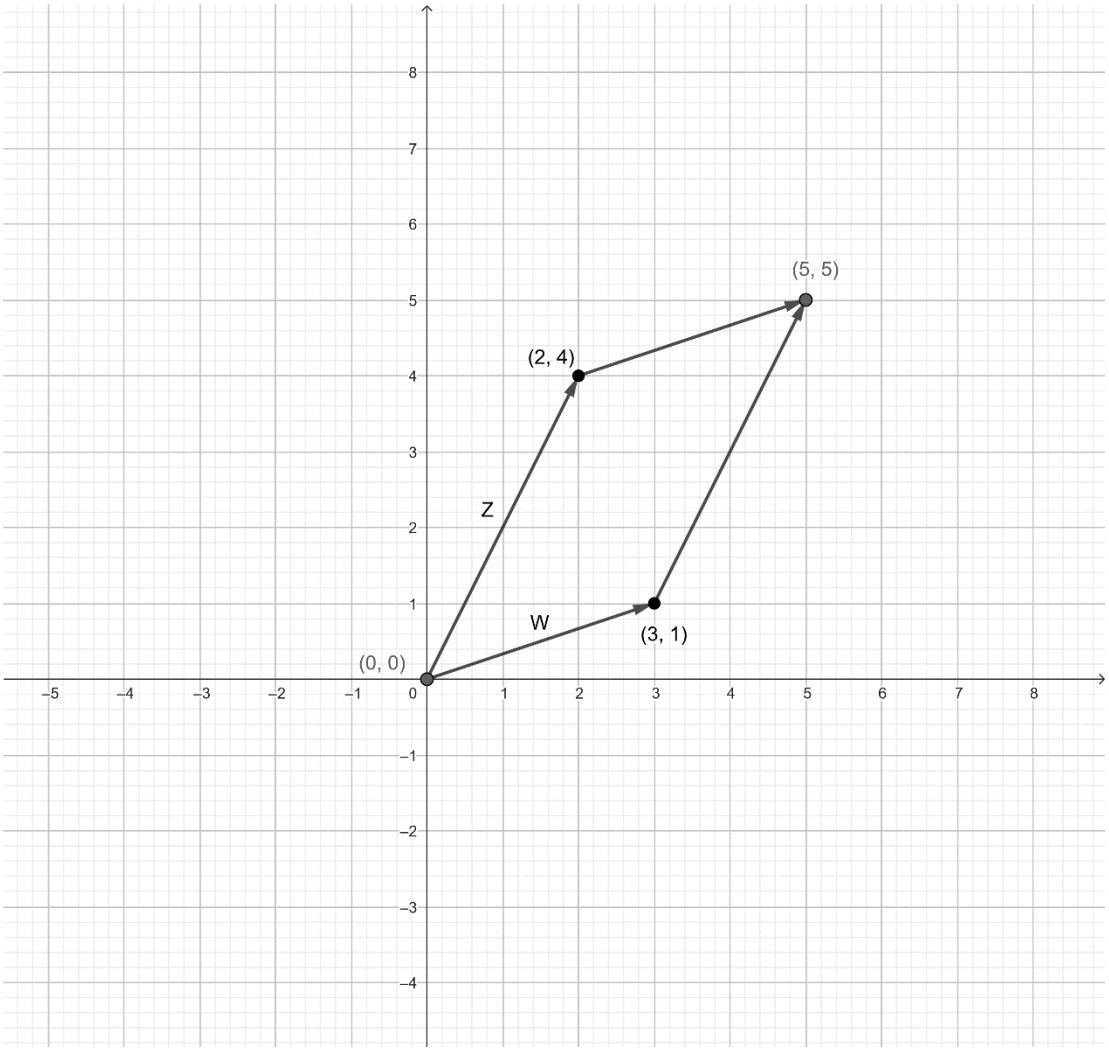


Figure 58. Three points depicted on the complex plane

Consider two complex numbers $z = a + ib$ and $w = c + id$. Addition is accomplished by separately adding the real and imaginary parts of each number (similar to vector addition):

$$z + w = (a + c) + i(b + d)$$

For example, the sum of $z = 2 + 4i$ and $w = 3 + i$ is $5 + 5i$. Figure 58 represents the addition of vectors (using the typical parallelogram approach for adding vectors). For a refresher on vector addition, see the YouTube video “Vector Addition” [85].



Subtraction of complex numbers is also done in a component-wise basis, e.g.,

$$(2 - 7i) - (4 + 5i) = (2 - 4) + (-7 - 5)i = -2 - 12i$$

We can derive a formula for the multiplication of two complex numbers using the distributive law, and then collecting the real and imaginary parts (and keeping in mind that $i^2 = -1$).

$$(a + bi)(c + di) = a(c + di) + bi(c + di) = ac + adi + bci - bd$$
$$= (ac - bd) + (ad + bc)i$$

In general, the commutative, associative and distributive laws hold for the set of complex numbers.

The **complex conjugate**, or simply the conjugate, of a complex number $z = a + bi$ is defined to be the complex number $a - bi$, and is denoted by $\bar{z}$. The conjugate of a complex number is the reflection of that number about the real axis.

Division of complex numbers can be accomplished by multiplying the numerator and denominator by the conjugate of the denominator.

$$\frac{a+bi}{c+di} = \frac{a+bi}{c+di} \cdot \frac{c-di}{c-di} = \frac{(ac+bd)+(bc-ad)i}{c^2+d^2}$$

Other than 0, every complex number has a multiplicative inverse. If $z = a + bi$ then its inverse is

$$z^{-1} = \frac{1}{a+bi} = \frac{1}{a+bi} \cdot \frac{a-bi}{a-bi} = \frac{a-bi}{a^2+b^2}$$

For example, the inverse of $3 + 4i$ is $\frac{3-4i}{25}$ which checks out as follows:

$$(3+4i) \cdot \frac{3-4i}{25} = \frac{1}{25}\big((9+16)+(12-12)i\big) = \frac{25}{25} = 1$$

. . .

The modulus of a complex number $z = a + ib$, denoted $|z|$, is its distance from the origin, which is easily computed using the Pythagorean theorem, i.e., $|z| = \sqrt{a^2+b^2}$ (see Figure 59). From basic trigonometric definitions, we also have that $a = |z| \cos\theta$ and $b = |z| \sin\theta$, and so, we have the following alternate form for z. This is known as the **polar form of a complex number**.

$$z = |z|(\cos\theta + i \sin\theta)$$

If $z = 0$, the coordinate θ is undefined. So, it is assumed that $z \neq 0$ whenever polar coordinates are used.

It is more common to use r instead of $|z|$ when using polar coordinates, i.e., $z = r(\cos\theta + i \sin\theta)$

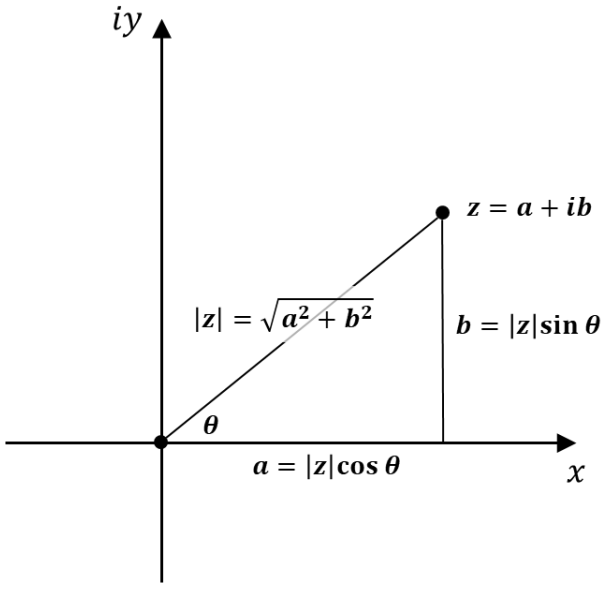


Figure 59. Modulus of a complex number

The parameter θ is referred to as the argument of z and is denoted $\arg z$. The value of $\arg z$ is not unique, i.e., it repeats every 2π radians. The **principal value of $\arg z$**, denoted by $Arg\ Z$, is usually taken to be $(-\pi, \pi]$ or $[0,2\pi)$. By selecting such an interval, we say that we have chosen a particular branch of $\arg z$ (this is referred to as the **principal branch**). We write $\arg z = Arg\ z + 2\pi n, n = 0, \pm 1, \pm 2, \ldots$ where $Arg\ z$ represents the selected principal branch.

If we take the principal branch $\arg z$ to be $(-\pi, \pi]$, then $\lim_{\epsilon \to 0} (\text{Arg}(-1 + \epsilon i)) = \pi$ and $\lim_{\epsilon \to 0} (\text{Arg}(-1 - \epsilon i)) = -\pi$. Thus, the negative real axis is a ray of discontinuity for $\arg z$. This ray of discontinuity is known as a **branch cut**, see Figure 60.

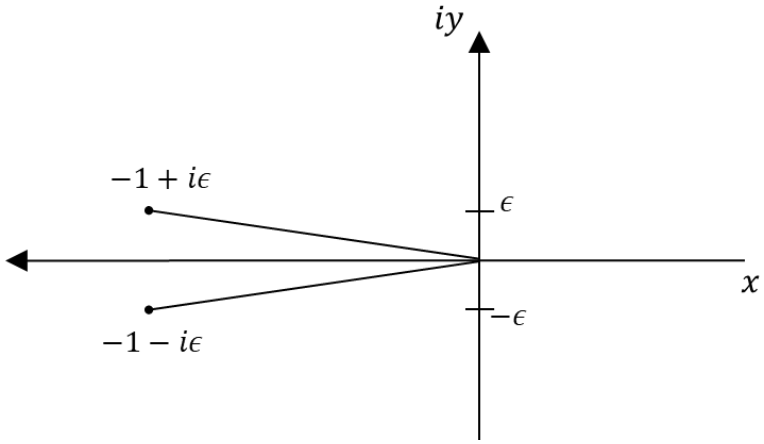


Figure 60. Branch cut along negative real axis

The polar form of a complex number can also be written in terms of the exponential function.

Theorem 112. $e^{i\theta} = cos\ \theta + i\ sin\ \theta$ for real θ.

Proof: Let

$$f(\theta) = \frac{cos\ \theta + i\ sin\ \theta}{e^{i\theta}} = e^{-i\theta}(cos\ \theta + i\ sin\ \theta)$$

Taking the derivative of $f(\theta)$ using the product rule, we get

$$f'(\theta) = e^{-i\theta}(-sin\ \theta + i\ cos\ \theta) - ie^{-i\theta}(cos\ \theta + i\ sin\ \theta) = 0$$

Since $f'(\theta) = 0$, it must be that $f(\theta)$ is constant. Further, we know that $f(0) = 1$ and so, $f(\theta) = 1$ for all real θ. ■

Using the above theorem and the previous derived polar form for a complex number gives us the famous **Euler's formula**:

$$z = re^{i\theta}, \qquad r = |z|, \qquad \theta = \arg(z)$$

This gives us the amazing relationship between e, i and π.

$$e^{i\pi} = -1$$

In addition to its many uses in complex analysis, we can also use Euler's formula to prove basic identities in trigonometry, e.g., the triple angle formula.

Consider $e^{3i\theta} = \cos 3\theta + i \sin 3\theta$ by Euler's formula.

On the other hand, we have the following expression

$$e^{3i\theta} = \left(e^{i\theta}\right)^3 = (cos\ \theta + i\ sin\ \theta)^3$$

$$= \cos^3(\theta) + 3i\cos^2(\theta)\sin(\theta) - 3\cos(\theta)\sin^2(\theta) - i\sin^3(\theta)$$

$$= \cos^3(\theta) - 3\cos(\theta)\sin^2(\theta) + i\,[3\cos^2(\theta)\sin(\theta) - \sin^3(\theta)]$$

$$= \cos^3(\theta) - 3\cos(\theta)\,(1 - \cos^2(\theta)) + i\,[3\,(1 - \sin^2(\theta))\sin(\theta) - \sin^3(\theta)]$$

$$= 4\cos^3(\theta) - 3\cos(\theta) + i\,[3\sin(\theta) - 4\sin^3(\theta)]$$

Equating the two expansions of $e^{3i\theta}$, we have

$$\cos(3\theta) = 4\cos^3(\theta) - 3\cos(\theta)$$

$$\sin(3\theta) = 3\sin(\theta) - 4\sin^3(\theta)$$

As an exercise, the reader is invited to derive the double angle formulas, i.e.,

$$\cos(2\theta) = \cos^2\theta - \sin^2\theta$$

$$\sin(2\theta) = 2\sin\theta\cos\theta$$

$$\ldots$$

We have the following relationship between a complex number $z = a + bi$ and its conjugate.

$$\sqrt{z\bar{z}} = \sqrt{(a + bi)(a - bi)} = \sqrt{a^2 + b^2} = |z|$$

This can also be written as $z\bar{z} = |z|^2$ which implies $\frac{1}{z} = \frac{\bar{z}}{|z|^2}$ when $z \neq 0$.

The identities in the following theorem are easy to prove by substituting $x + iy$ for z and $u + iv$ for w.

Theorem 113. For complex numbers z and w, we have the following identities:

$$Re(z) = \frac{1}{2}(z + \bar{z}), \qquad Im(z) = \frac{1}{2}(z - \bar{z}), \qquad |z| \geq |Re(z)|$$

$$\bar{\bar{z}} = z, \qquad |\bar{z}| = |z|, \qquad |zw| = |z|\,|w|, \qquad \left|\frac{z}{w}\right| = \frac{|z|}{|w|}$$

$$\overline{z \pm w} = \bar{z} \pm \overline{w}, \qquad \overline{zw} = \bar{z} \cdot \overline{w}, \qquad \overline{\left(\frac{z}{w}\right)} = \frac{\bar{z}}{\overline{w}}$$

$$(\bar{z})^k = \overline{z^k}, \qquad \overline{az^k} = a\,\bar{z}^k \text{ for real constant } a$$

Using some of the above identities, we can prove the **triangle inequality for complex numbers**, i.e., $|z + w| \leq |z| + |w|$.

$$|z + w|^2 = (z + w)(\overline{z + w}) = (z + w)(\bar{z} + \overline{w}) = z\bar{z} + z\overline{w} + \overline{w}z + w\overline{w}$$

$$= |z|^2 + z\overline{w} + \overline{z\overline{w}} + |w|^2 = |z|^2 + 2Re(z\overline{w}) + |w|^2$$

$$\leq |z|^2 + 2|Re(z\overline{w})| + |w|^2 \leq |z|^2 + 2|z\overline{w}| + |w|^2 = |z|^2 + 2|z||w| + |w|^2 = (|z| + |w|)^2$$

Taking the square root on both sides of the above, we get $|z + w| \leq |z| + |w|$. Since $|w| = |-w|$, we can replace w with $-w$ in the triangle inequality to get $|z - w| \leq |z| + |w|$.

We can derive another useful inequality from the triangle inequality as follows:

$$|z| = |(z + w) - w| \leq |z + w| + |-w| = |z + w| + |w|$$

This gives us $|z + w| \geq |z| - |w|$. By symmetry, we also have $|z + w| \geq |w| - |z|$. We can combine the two results into the following inequality:

$$|z + w| \geq \big||z| - |w|\big|$$

We can replace w with $-w$ in the above to get yet another inequality, i.e.,

$$|z - w| \geq \big||z| - |w|\big|$$

By way of mathematical induction, we can extend the triangle inequality for the case of n complex numbers, i.e.,

$$|z_1 + z_2 + \cdots + z_n| \leq |z_1| + |z_2| + \cdots + |z_n|$$

. . .

The distance between two complex numbers is defined in exactly the same way as the Euclidean distance between points in $\mathbb{R}^2$. For $z = a + bi$ and $w = c + di$, the distance between z and w is defined as

$$d(z, w) = |z - w| = \sqrt{(a - c)^2 + (b - d)^2}$$

4.3 Roots of Complex Numbers

Consider the solutions to $z^n = 1$ for complex number z and natural numbers $n = 1,2,3, \ldots$ (denoted by the symbol $\mathbb{N}$).

For $n = 1$, the only solution is $z = 1$, and for $n = 2$, the only solutions are 1 and -1. So far, this is the same as the real number case.

For $n = 3$, we clearly have 1 as a solution, but there are two more solutions. To see this, we represent 1 as $e^{i(2\pi m)}, m = 0,1,2, \ldots$

So, $z^3 = 1 = e^{i(2\pi m)}, m = 0,1,2, \ldots$ which implies

$$z = e^{i\left(\frac{2m\pi}{3}\right)}, m = 0,1,2, \ldots$$

For $m = 0$, we get the solution $e^0 = 1$

For $m = 1$, we get the solution $e^{i\left(\frac{2\pi}{3}\right)} = \cos\left(\frac{2\pi}{3}\right) + i\sin\left(\frac{2\pi}{3}\right) = -\frac{1}{2} + \frac{i\sqrt{3}}{2}$

For $m = 2$, we get the solution $e^{i\left(\frac{4\pi}{3}\right)} = \cos\left(\frac{4\pi}{3}\right) + i\sin\left(\frac{4\pi}{3}\right) = -\frac{1}{2} - \frac{i\sqrt{3}}{2}$

For $m = 3$, we are back to the solution 1 and the pattern repeats from here. So, we have three solutions.

We can use the same technique to solve for the solutions of $z^n = 1$ in general.

Let $z^n = 1 = e^{2\pi m}, m = 0,1,2, \ldots$ which implies

$$z = e^{i\left(\frac{2m\pi}{n}\right)} = \cos\left(\frac{2m\pi}{n}\right) + i\sin\left(\frac{2m\pi}{n}\right), \qquad m = 0,1,2, \ldots, n - 1$$

The pattern starts to repeat when $m = n$ and so, we have exactly, n solutions. The solutions to $z^n = 1$ are known as the **roots of unity**.

Figure 61 depicts the five 5th roots of unity. Each root lies on the unit circle. The roots are $\frac{2\pi}{5}$ radians (72°) apart.

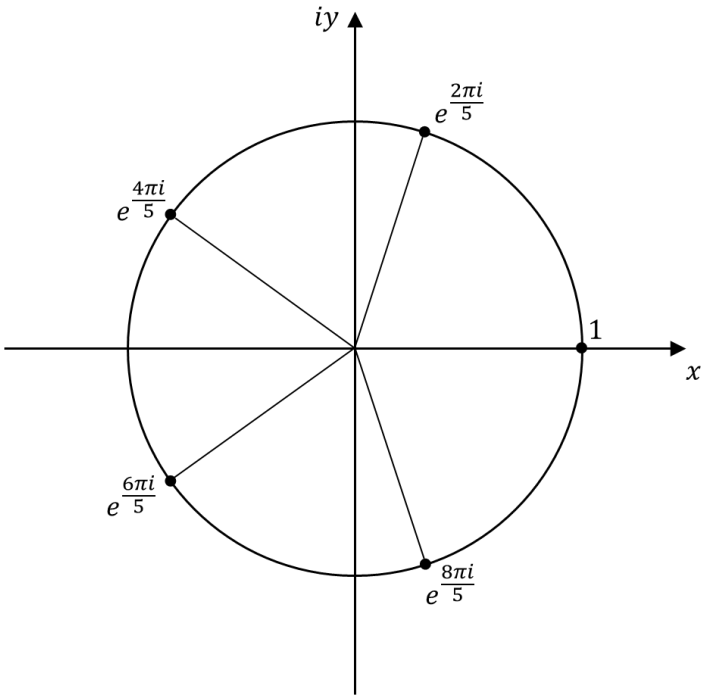


Figure 61. Fifth roots of unity

It is also possible to find the n^{th} roots of any complex number $z_0 = re^{i\phi}$ where ϕ is the principal value for the argument of z_0, i.e., $\phi \in (-\pi, \pi]$.

We need to solve for solutions to $z^n = z_0 = re^{i(\phi+2\pi m)},\ m = 0,1,2, \ldots$

Taking the n^{th} root on both sides of the above equation yields

$$z = \sqrt[n]{r}\, e^{i\left(\frac{\phi}{n}+\frac{2\pi m}{n}\right)} = \sqrt[n]{r}\left[\cos\left(\frac{\phi}{n}+\frac{2\pi m}{n}\right) + i\sin\left(\frac{\phi}{n}+\frac{2\pi m}{n}\right)\right], \qquad m = 0,1,2, \ldots, n-1$$

For $m = n, n+1, \ldots$ the roots repeat.

Figure 62 depicts a geometric representation of the 2nd to 6th roots of a general complex number $z = re^{i\phi}$.

The figure is from the Wikipedia article “Root of unity” [86].

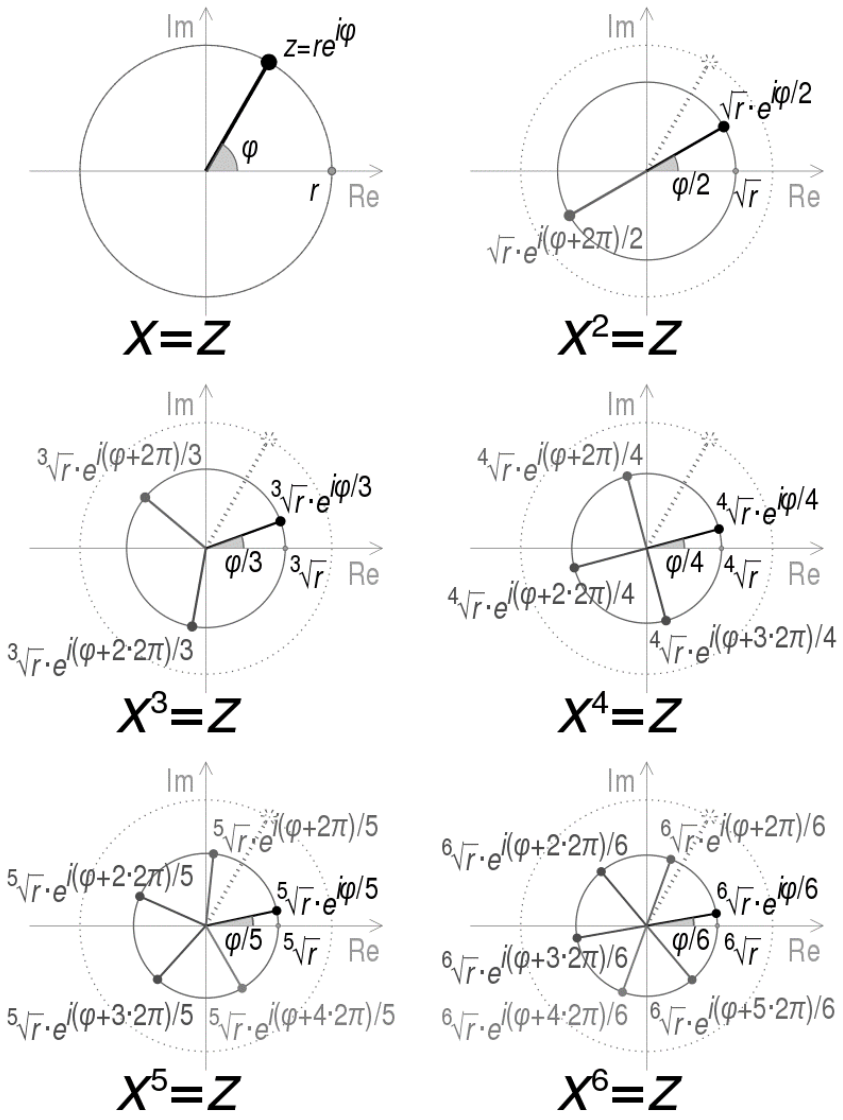


Figure 62. Representation of the 2nd to 6th root of a complex number

An n^{th} degree single-variable complex polynomial has the form

$$z_n z^n + z_{n-1} z^{n-1} + \cdots + z_1 z + z_0$$

where z is a complex variable, and the z_i terms are complex constants.

In general, every n^{th} degree single-variable, complex polynomial has n roots (some of which may be repeated). This result is stated in the following theorem.

Theorem 114. (Fundamental Theorem of Algebra) The following statements are equivalent:

- Every non-constant single-variable complex polynomial has at least one complex root.
- Every single-variable complex polynomial with complex coefficients can be factorized as $a(z-c_1)(z-c_2)\dots(z-c_n)$, where a is a complex number, and $c_i, i = 1,2,\dots,n$ are the roots of the given polynomial. Some of the c_i terms may be repeated.

Proof: Concerning the first statement, a proof can be found in the Wikipedia article "Fundamental theorem of algebra" [87] for several different proofs.

Assume the first statement is true and that an n^{th} degree complex polynomial $p_n(z)$ has complex root z_1. Using polynomial long division, we can write

$$p_n(z) = (z - z_1)\, p_{n-1}(z) + c$$

where $p_{n-1}(z)$ is of degree $n-1$ and c is a complex constant. However, since c_1 is a root of $p_n(z)$, we have $0 = p_n(c_1) = (c_1 - c_1)p_{n-1}(c_1) + c = c$. So, $p_n(z) = (z - c_1)\, p_{n-1}(z)$. The first statement of the theorem tells us that $p_{n-1}(z)$ has at least one complex root (say c_2). Factor out the root to get $p_n(z) = (z - c_1)(z - c_2)\, p_{n-2}(z)$ where $p_{n-2}(z)$ is of degree $n-2$. Continuing the process, we can eventually write $p_n(z) = a(z-c_1)(z-c_2)\dots(z-c_n)$ where $c_1, c_2, \dots, c_n$ are the roots of $p_n(z)$ and $a \in \mathbb{C}$. Some of the roots could be repeated. Thus, a polynomial is completely determined by its roots up to a constant multiple. ■

4.4 Relation to metric spaces and topological spaces

Using the metric $d(z,w) = \sqrt{|z-w|}$, the complex plane $\mathbb{C}$ forms a metric space.

The complex plane is also a topological space with its open sets being those induced by the metric space.

As topological spaces, $\mathbb{R}^2$ and $\mathbb{C}$ are homeomorphic.

By Theorem 103, $\mathbb{C}$ is T_4, i.e., normal Hausdorff.

$\mathbb{C}$ is a second countable space, but it is not compact.

$\mathbb{C}$ is a connected space.

4.5 Complex Functions

4.5.1 Overview

Given a set $S \subseteq \mathbb{C}$, **complex function** f is a mapping from S into $\mathbb{C}$, which we represent as $f: S \to \mathbb{C}$. For each $z \in S$, there is a corresponding $w = f(z)$. For example, if $f(z) = z^2$ then $f(1+i) = (1+i)^2 = 1 + 2i - 1 = 2i$. The set S is known as the **domain** of the function f, and the set $f(S) = \{f(z): z \in S\} \subseteq \mathbb{C}$ is know as the **range or codomain** of f.

Warning: Various books and articles on complex analysis often use the term "domain" in two different ways. One way is what we just stated above, i.e., as the set of values over which a function is defined. In addition, a domain is defined as an open connected set. Key point is that the

domain of a complex function is not necessarily an open connected set. Yes, this is confusing, but it's not my idea.

If for each $z \in S, f(z)$ maps to only one value in the range, we say that f is a **single-valued function**. For example, $f(z) = z^3$ is a single=valued function. If $f(z)$ maps to more than one value in the range for each z in the domain, then f is said to be a **multi-valued function**. The function $f(z) = \sqrt{z}$ is a multi-valued, e.g., $f(-1) = \sqrt{-1} = \pm i$. A multi-valued function can be viewed (represented) as a collection of single-valued functions, each member of which is called a **branch** of the function. Typically, we consider one particular member as a **principal branch** of the multiple-valued function.

Let $w = u + iv$ be the value of a function f at $z = x + iy$, i.e., $f(x + iy) = u + iv$. Since of the real numbers u and v depend on the real variables x and y, it follows that $f(z)$ can be expressed in terms of a pair of real-valued functions of the real variables x and y, i.e.,

$$f(z) = u(x, y) + i\, v(x, y)$$

If we use the polar coordinates r and θ, instead of x and y, then we can represent f as

$$f(z) = u(r, \theta) + i\, v(r, \theta)$$

For example, take $f(z) = z^3$.

$$f(z) = f(x + iy) = (x + iy)^3 = x^3 + 3ix^2y - 3xy^2 - iy^3$$

$$= (x^3 - 3xy^2) + i\,(3x^2y - y^3)$$

Using polar coordinates, we have

$$f\left(re^{i\theta}\right) = r^3 e^{3i\theta} = r^3(\cos 3\theta + i \sin 3\theta)$$
$$= r^3 \cos 3\theta + i\, r^3 \sin 3\theta$$

4.5.2 Plotting

Plotting of complex functions requires four dimensions since such mappings are from one 2-dimensional space (the domain of a function) to another (the codomain of the function). There are various techniques that attempt to get around this issue. The simplest technique is to plot only part of the domain. For example, consider the function $f(z) = z^2$. We can get an idea of the mapping by plotting horizontal and vertical lines from the domain to the codomain. Horizontal lines are given by the equations $y = c$ where c is a constant, and the vertical lines are given by the equations $x = c$ where c is a constant.

Expanding the mapping in terms of x and y, we have

$$f(z) = f(x + iy) = (x + iy)^2 = (x^2 - y^2) + i\,(2xy)$$

When $y = c$, the mapping becomes

$$f(x + ic) = (x^2 - c^2) + i\,(2cx)$$

Let $u = x^2 - c^2$ and $v = 2cx$. Solving for x in term of v in the latter equation, we get $x = \frac{v}{2c}$.

Plugging this into the equation for u gives us $u = \left(\frac{v}{2c}\right)^2 - c^2$ which is a parabola opening to the right. In the case $c = 0$, we get the positive horizontal ray defined by the points $(u, v) = (x^2, 0)$.

When $x = c$, the mapping becomes

$$f(c + iy) = (c^2 - y^2) + i\,(2cy)$$

Let $u = c^2 - y^2$ and $v = 2cy$. Solving for y in terms of v in the latter equation, we get $y = \frac{v}{2c}$.

Plugging this into the equation for u gives us $u = c^2 - \left(\frac{v}{2c}\right)^2$ which is a parabola opening to the left. In the case $c = 0$, we get the negative horizontal ray defined by the points $(u, v) = (-y^2, 0)$.

One can try out some mapping of lines using the online tool at https://tobylam.xyz/plotter/. Some online applications will show how a grid of vertical and horizontal lines in the domain get mapped to the codomain, e.g., try entering $f(z) = \frac{1}{z}, f(z) = e^z$ and $f(z) = \log(z)$ at www.wolframalpha.com and scroll down to the section entitled "Complex map". In addition, the YouTube video from Michael Penn [88] provides some additional examples.

Another way to visualize a complex function is to consider translations, rotations and reflections.

- For example, $f(z) = z + (1 + i)$ translates points in the domain one unit up and one unit right in the codomain.
- In terms of rotations, consider $f(z) = z^3$ in terms of polar coordinates, i.e., $f\left(re^{i\theta}\right) = r^3 e^{i3\theta}$. The argument of z get mapped to r^3. For $r > 1$, the argument gets larger in the codomain, and for $r < 1$, the argument gets smaller. The angle is multiplied by 3 in the mapping. If you try mapping a grid in the domain via one of the online applications mentioned above, you will see that the situation is quite complicated.
- In terms of reflections, a simple example is $f(z) = \bar{z}$. This function exchanges the upper and lower half-planes.

Other more complex mapping techniques involve the use of colors to indicate the argument and/or modulus. The following vides provide summaries of such techniques:

- What does a complex function look like? [89]
- The 5 ways to visualize complex functions [90].

4.5.3 Limits

For real-valued functions, an open ball around a point is a 1-dimensional set, i.e., (0,2) is an open ball about the point $1 \in \mathbb{R}$. Thus, for the limit about a point to exist, one only needs to determine that the limit is the same as one approaches from the right and the left. For example, consider the function

$$f(x) = \begin{cases} 1, & x > 1 \\ 0, & x = 1 \\ -1, & x < 1 \end{cases}$$

The limit as $x \to 0$ does not exist since the limit as one approaches from the left is -1 which does not equal the limit as one approaches from the right which equals 1.

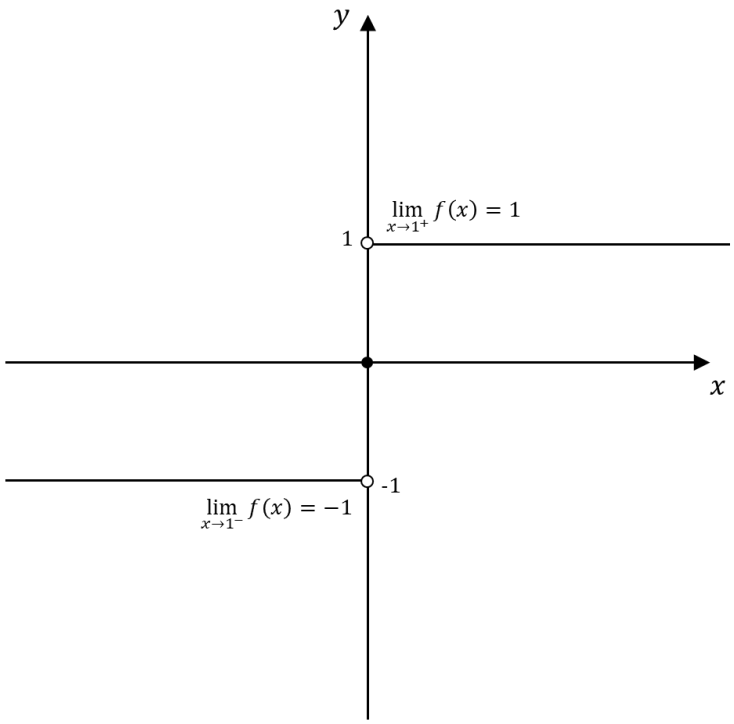


Figure 63. Example – limit of a real-valued function

For complex functions, an open ball surrounding a point is 2-dimensional (the interior of a disk to be exact). For a limit of a function $f(z)$ to exist at a point z_0, the function must approach the same value from all directions leading to z_0.

Consider the complex function

$$f(z) = \frac{z}{\bar{z}} = \frac{x + iy}{x - iy}$$

We will compute the limit from two different directions to show that the limit does not exist at $z = 0$. First, take the limit along the positive real axis (holding $y = 0$)

$$\lim_{x \to 0^+} f(z) = \lim_{x \to 0^+} \frac{x + i0}{x - i0} = 1$$

Next, take the limit along the positive imaginary axis (holding x=0)

$$\lim_{y \to 0^+} f(z) = \lim_{y \to 0^+} \frac{0 + iy}{0 - iy} = -1$$

. . .

In terms of notation, we write the **limit** of $f(z)$ equals w_0 as z approaches z_0 as

$$\lim_{z \to z_0} f(z) = w_0$$

This means that the point $w = f(z)$ can be made arbitrarily close to w_0 if we chose the point z close enough to z_0 but distinct from it. Formally, the limit of $f(z)$ equals w_0 as z approaches z_0 if for each positive real number ε, there is a positive real number δ such that $|f(z) - w_0| < \epsilon$ whenever $0 < |z - z_0| < \delta$.

Warning: This is not the same concept as a limit point that we defined for metric spaces.

Theorem 115. When a limit of a function exists at a point, it is unique.

Proof: Assume that

$$\lim_{z \to z_0} f(z) = w_0, \qquad \lim_{z \to z_0} f(z) = w_1$$

For every $\frac{\epsilon}{2} > 0$, there exists $\delta_0 > 0$ and $\delta_1 > 0$ such that

$$|f(z) - w_0| < \epsilon \text{ whenever } 0 < |z - z_0| < \delta_0, \text{ and}$$

$$|f(z) - w_1| < \epsilon \text{ whenever } 0 < |z - z_0| < \delta_1$$

If we choose $\delta = \min(\delta_0, \delta_1)$, then whenever $0 < |z - z_0| < \delta$, we have the following (using the triangle inequality)

$$|w_0 - w_1| \leq |f(z) - w_0| + |f(z) - w_1| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

Since we can choose ϵ arbitrarily small, we must conclude the $w_0 = w_1$. ■

As an example, we will show that $\lim_{z \to z_0} z^2 = z_0^2$.

For any $\epsilon > 0$, we need to find $\delta > 0$ such that $\left|z^2 - z_0^2\right| < \epsilon$ whenever $0 < |z - z_0| < \delta$. If $\delta \leq 1$, then $0 < |z - z_0| < \delta$ implies

$$\left|z^2 - z_0^2\right| = |z - z_0|\,|z + z_0| < \delta\,|z - z_0 + 2z_0| < \delta(|z - z_0| + |2z_0|) < \delta(1 + 2|z_0|)$$

So, if we take $\delta = \min(1, \frac{\epsilon}{1+2|z_0|})$, then $\left|z^2 - z_0^2\right| < \epsilon$ whenever $|z - z_0| < \delta$.

As a second example, we will show that $\lim_{z \to 1} i\bar{z} = i$. First, notice that

$$|f(z) - i| = |i\bar{z} - i| = |i|\,|\bar{z} - i| = 1 \cdot |z - 1| = |z - 1|$$

Thus, for any $\epsilon > 0$,

$$|f(z) - i| < \epsilon, \text{ whenever } 0 < |z - 1| < \epsilon = \delta$$

The limit of a complex function is directly related to the associated real-valued functions of two variables that represent the real and imaginary parts of the complex function.

Theorem 116. Given $f(z) = u(x, y + i\, v(x, y)$ where $z = x + iy$ and u and v are real-valued functions of two variables. We have that

$$\lim_{z \to z_0} f(z) = w_0 \text{ if and only if}$$

$$\lim_{(x,y)\to(x_0,y_0)} u(x,y) = u_0 \text{ and } \lim_{(x,y)\to(x_0,y_0)} v(x,y) = v_0$$

$$\text{where } z_0 = x_0 + iy_0 \text{ and } w_0 = u_0 + iv_0.$$

Proof: See Theorem 1 of Section 16 in the book by Brown and Churchill [91]. ■

Theorem 116 and the theorems for limits of real-valued functions (from basic calculus) can be used to prove the following theorem concerning the limits of complex functions.

Theorem 117. If $\lim_{z \to z_0} f(z) = w_1$ and $\lim_{z \to z_0} g(z) = w_2$, then

$$\lim_{z \to z_0} af(z) = aw_1, \qquad a \in \mathbb{C}$$

$$\lim_{z \to z_0} f(z) \pm g(z) = w_1 \pm w_2$$

$$\lim_{z \to z_0} f(z) \cdot g(z) = w_1 \cdot w_2$$

$$\lim_{z \to z_0} \frac{f(z)}{g(z)} = \frac{w_1}{w_2}, w_2 \neq 0$$

Theorem 118. The limit exists at every point $z \in \mathbb{C}$ for the complex polynomial $f(x) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$.

Proof: We first note that for the constant function $f(z) = c$ and the identity function $f(z) = z$, the limit exists at every point z. This follows directly from the definition of a limit.

The theorem follows by multiple applications of Theorem 117. ■

. . .

It is possible to define limits of complex functions (and sequences) as $z \to \infty$ but we first need to add the so-called "point at infinity" to the complex plane. From the Wikipedia article "Riemann sphere" [92]:

> In mathematics, the Riemann sphere, named after Bernhard Riemann, is a model of the extended complex plane (also called the closed complex plane): the complex plane plus one point at infinity. This extended plane represents the extended complex numbers, that is, the complex numbers plus a value ∞ for infinity. With the Riemann model, the point ∞ is near to very large numbers, just as the point 0 is near to very small numbers.

Figure 64 (taken from "Riemann sphere" [92]) shows the stereographic projection of the unit sphere onto the extended complex plane. The sphere is centered at the origin and is divided into

two hemispheres by the complex plane. There is a one-to-one mapping between the points on the sphere and the extended complex plane. The top point of the sphere, labelled as $P(\infty)$, is mapped to the point at infinity of the complex plane. For a point A in the extended complex plane and outside of the sphere, we draw a line from $P(\infty)$ to A. The point of intersection on the upper hemisphere, shown as $\alpha = P(A)$, is the mapping from the sphere to point A. For points within the sphere (e.g., B), we do the same procedure, except that the point of intersection $\beta = P(B)$ is on the bottom hemisphere.

There are some other items labelled in the figure, but they are not relevant to the discussion here.

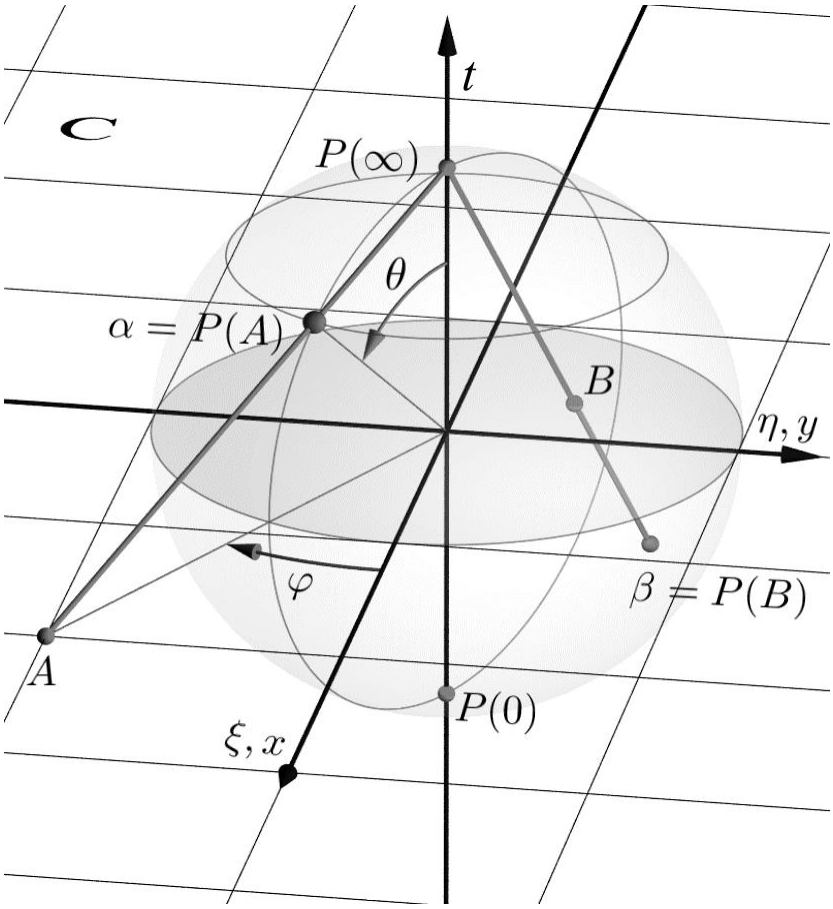


Figure 64. Stereographic projection of unit sphere onto extended complex plane

There are three cases with regard to the extended complex plane and limits.

Case 1: $\lim_{z \to z_0} f(z) = \infty$ means that for each $\epsilon > 0$, there exists $\delta > 0$ such that

$$|f(z)| > \frac{1}{\epsilon} \text{ whenever } 0 < |z - z_0| < \delta$$

This case is equivalent to $\lim_{z \to z_0} \frac{1}{f(z)} = 0$.

For example, $\lim_{z \to i} \frac{z+i}{z-i} = \infty$ since, using the equivalent expression, we have

$$\lim_{z\to i}\frac{z-i}{z+i}=0$$

Case 2: : $\lim_{z\to\infty} f(z)=w_0$ means that for each $\epsilon>0$, there exists $\delta>0$ such that

$$|f(z)-w_0|<\epsilon \text{ whenever } |z|>\frac{1}{\delta}$$

This case is equivalent to $\lim_{z\to 0} f\left(\frac{1}{z}\right)=w_0$.

For example, $\lim_{z\to\infty}\frac{3z+2i}{5z-i}=\frac{3}{5}$ since, using the equivalent expression, we have

$$\lim_{z\to 0}\frac{\frac{3}{z}+2i}{\frac{5}{z}-i}=\lim_{z\to 0}\frac{3+2iz}{5-iz}=\frac{3}{5}$$

Case 3: $\lim_{z\to\infty} f(z)=\infty$ means that for each $\epsilon>0$, there exists $\delta>0$ such that

$$|f(z)|>\frac{1}{\epsilon} \text{ whenever } |z|>\frac{1}{\delta}$$

This case is equivalent to $\lim_{z\to 0}\frac{1}{f\left(\frac{1}{z}\right)}=0$.

For example, $\lim_{z\to\infty}\frac{z^4+z}{z^3+1}=0$ since, using the equivalent expression, we have

$$\lim_{z\to 0}\frac{\left(\frac{1}{z}\right)^3+1}{\left(\frac{1}{z}\right)^4+\frac{1}{z}}=\lim_{z\to 0}\frac{z+z^4}{1+z^3}=0$$

4.5.4 Continuity

Since the complex plane is a topological space with the open sets generated by the associated metric (i.e., the modulus that we defined earlier), we can use any of the three equivalent definitions of continuity from Section 3.4.7. However, we will usually use the first definition, i.e., a complex function $f(z)$ is continuous at a point z_0 if

$$\lim_{z\to z_0} f(z)=f(z_0)$$

The above equations implies that the limit exists and $f(z_0)$ is defined.

The various theorems for continuous functions in topological spaces (in Section 3.5.4) also hold for complex functions, e.g., the composition of two continuous complex functions is also continuous.

Every complex polynomial, i.e., a complex function of the form $f(x)=a_nz^n+a_{n-1}z^{n-1}+\cdots+a_1z+a_0$, is continuous at every point. Since such functions are defined at every point, and the associated limit approaching a given point z_0 clearly equals $f(z_0)$.

It follows from Theorem 116 that a complex function $f(z) = f(x+iy) = u(x,y) + iv(x,y)$, with $u(x,y)$ and $v(x,y)$ being real-valued functions, is continuous if and only if $u(x,y)$ and $v(x,y)$ are continuous.

4.5.5 Derivatives

Derivatives for complex functions are defined in the same way as for real-valued functions. The complex function $f(z)$, whose domain contains the open all $B(z_0;\epsilon)$, is differentiable at z_0 if the following limit exists

$$f'(z_0) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}$$

Alternately (and similar to what is done for real-valued functions in differential calculus), we can define the complex variable $\Delta z = z - z_0, z \neq z_0$ and rewrite the derivative as

$$f'(z_0) = \lim_{\Delta z \to 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z}$$

As an example, consider the function $f(z) = cz^n$ for $c \in \mathbb{C}$ and $n \in \mathbb{N}$. Using the alternate definition of a derivative at any point z and the binomial theorem, we have

$$f'(z) = \lim_{\Delta z \to 0} \frac{f(z+\Delta z) - f(z)}{\Delta z} = \lim_{\Delta z \to 0} \frac{c(z+\Delta z)^n - cz^n}{\Delta z}$$

$$= c\left[\lim_{\Delta z \to 0} \frac{z^n + \binom{n}{1} z^{n-1}\Delta z + \binom{n}{2} z^{n-2}(\Delta z)^2 + \cdots + \binom{n}{n-1} z(\Delta z)^{n-1} + (\Delta z)^n - z^n}{\Delta z}\right]$$

$$= c\left[\lim_{\Delta z \to 0} \binom{n}{1} z^{n-1} + \binom{n}{2} z^{n-2}\Delta z + \cdots + \binom{n}{n-1} z(\Delta z)^{n-2} + (\Delta z)^{n-1}\right] = cnz^{n-1}$$

As we saw with basic limits, for a limit to exist, it must be the same for all directions approaching the point in question (recall the example concerning $z/\bar{z}$). As the derivative is defined as a limit, we have the same requirement. For example, consider the derivative of $f(z) = \bar{z}$ at any point z. First, we let $\Delta z = (\Delta x, \Delta y)$ approach 0 along the positive imaginary axis, i.e.,

$$\lim_{\Delta y \to 0, \Delta x = 0} \frac{f(i(y+\Delta y)) - f(iy)}{i\Delta y} = \lim_{\Delta y \to 0, \Delta x = 0} \frac{-i(y+\Delta y) - (-iy)}{i\Delta y} = -1$$

We then compute the derivative as we approach 0 along the positive real axis, i.e.,

$$\lim_{\Delta x \to 0, \Delta y = 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} = \lim_{\Delta x \to 0, \Delta y = 0} \frac{(x+\Delta x) - (x)}{\Delta x} = 1$$

So, since the limit is not the same as Δz approach 0 from all directions, the derivative of $\bar{z}$ and that is true for all points $z \in \mathbb{C}$.

It is possible for a complex function to have a derivative at only one point. For example, consider the function $f(z) = |z|^2$. Using the alternate definition of a derivative, we have

$$f'(z) = \lim_{\Delta z \to 0} \frac{f(z+\Delta z) - f(z)}{\Delta z} = \lim_{\Delta z \to 0} \frac{|z+\Delta z|^2 - |z|^2}{\Delta z} = \lim_{\Delta z \to 0} \frac{(z+\Delta z)(\overline{z+\Delta z}) - z\bar{z}}{\Delta z}$$

$$= \lim_{\Delta z \to 0} \frac{z\bar{z} + \bar{z}\,\Delta z + z\,\overline{\Delta z} + (\Delta z)^2 - z\bar{z}}{\Delta z} = \lim_{\Delta z \to 0} \bar{z} + \frac{z\,\overline{\Delta z}}{\Delta z} + \Delta z$$

Letting Δz approach 0 along the positive imaginary axis, we have that $\overline{\Delta z} = -\Delta z$ and the above limit evaluates to $\bar{z} - z$.

Letting Δz approach 0 along the positive real axis, we have that $\overline{\Delta z} = \Delta z$ and the above limit evaluates to $\bar{z} + z$.

For the derivative to exist, it is necessary for $\bar{z} - z = \bar{z} + z$ which is only true when $z = 0$. We can check that the derivative does exist at $z = 0$ using the definition of derivative, i.e.,

$$f'(0) = \lim_{z \to 0} \frac{f(z) - f(0)}{z - 0} = \lim_{z \to 0} \frac{|z|^2}{z} = \lim_{z \to 0} \frac{z\,\bar{z}}{z} = \lim_{z \to 0} \bar{z} = 0$$

Interestingly, $f(z) = |z|^2$ is continuous everywhere. To see this, divide the function into its real and imaginary parts, i.e.,

$$f(x + iy) = (x^2 + y^2) + i \cdot 0$$

Since the real and imaginary parts of $f(z)$ are continuous at each point $z \in \mathbb{C}$, f itself is continuous for all $z \in \mathbb{C}$. So, the continuity of a function at a point does not imply the existence of a derivative there. On the other hand, we have the following theorem.

Theorem 119. The existence of the derivative of a function at a point implies the continuity of the function at that point.

Proof: Assume that $f'(z)$ exists, i.e.,

$$f'(z_0) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}$$

We have the following

$$\lim_{z \to z_0} [f(z) - f(z_0)] = \lim_{z \to z_0} \left[\frac{f(z) - f(z_0)}{z - z_0}\right] \lim_{z \to z_0} [z - z_0] = f'(z_0) \cdot 0 = 0$$

So, $\lim_{z \to z_0} f(z) = f(z_0)$, i.e., the limit of $f(z)$ exists at z_0. ■

. . .

The formulas for the sum, difference, product, quotient and composition of complex functions are the same as those for real functions, i.e.,

$$(cf)'(z) = cf'(z), \qquad c \in \mathbb{C}$$

$$(f \pm g)'(z) = f'(z) \pm g'(z)$$

$$(fg)'(z) = f(z)g'(z) + f'(z)g(z)$$

$$\left(\frac{f}{g}\right)'(z) = \frac{f'(z)g(z) - f(z)g'(z)}{g^2(z)}$$

The **chain rule** for complex functions is again the same as that for real functions. If we let $F(z) = f\big(g(z)\big)$, then $F'(z) = f'\big(g(z)\big)g'(z)$. Alternately, we can write $w = g(z)$ and $W = f(w) = f\big(g(z)\big) = F(z)$ and the chain rule is then written as

$$\frac{dW}{dz} = \frac{dW}{dw} \cdot \frac{dw}{dz}$$

For example, consider $F(z) = (7z^2 - 3z + 5)^4$. If we let $g(z) = 7z^2 - 3z + 5$ and $f(z) = z^4$, then $F'(z) = f'\big(g(z)\big)g'(z) = 4\big(g(z)\big)^3(14z - 3) = 4(7z^2 - 3z + 5)^3(14z - 3)$.

. . .

The **Cauchy-Reimann theorem** [93] embodies a method for testing whether or not a complex function is differentiable. The Cauchy-Reimann theorem is as follows:

Given a complex function separated into its real and imaginary parts, i.e., $f(z) = u(x, y) + iv(x, y)$. If $f'(z)$ exists at a point $z_0 = x_0 + iy_0$ then the following equations (known as the Cauchy-Reimann equations) hold true at (x_0, y_0):

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$$

The above entities are partial derivatives, e.g., $\frac{\partial u}{\partial x}$ is the partial derivative of $u(x, y)$ with respect to x. The alternate notation u_x is often used instead of $\frac{\partial u}{\partial x}$ (and similar for the other partial derivatives).

The converse of the Cauchy-Reimann theorem is also true, i.e., if the partial derivatives of the real and imaginary parts of a complex function $f(z)$ exists, are continuous, and satisfy the Cauchy-Riemann equations at a point $z_0 = x_0 + iy_0$, then $f'(z_0) = u_x(x_0, y_0) + iv_x(x_0, y_0)$.

For a proof of the Cauchy-Reimann theorem and its converse, see Section 2.4 of "Fundamentals of complex analysis with applications to engineering and science" [95].

Consider our previous example of $f(z) = \bar{z} = x + (-y)i$. We have that

$$u_x = 1 \neq -1 = v_y, \qquad u_y = 0 = v_x$$

The Cauchy-Riemann equations do not hold, and thus, $f(z)$ is nowhere differentiable.

We can use the converse of the Cauchy -Riemann theorem to show that $f(z) = e^z$ is differentiable for all $z \in \mathbb{C}$. First, we expand $f(z)$ using Euler's formula, i.e.,

$$e^z = e^{x+iy} = e^x e^{iy} = e^x \cos y + i\, e^x \sin y$$

Letting $u(x, y) = e^x \cos y$ and $v(x, y) = e^x \sin y$, we have

$$u_x = e^x \cos y = v_y$$

$$u_y = -e^x \sin y = -v_x$$

So, the Cauchy-Riemann equations hold true. In addition, the partial derivatives of u and v exists and are continuous for all $x + iy \in \mathbb{C}$. Thus, $f'(z)$ exists for all $z \in \mathbb{C}$ and is equal to

$$u_x + iv_x = e^x \cos y + ie^x \sin y = e^x(\cos y + i \sin y) = e^x e^{iy} = e^{x+iy} = e^z$$

. . .

There is also a polar coordinates form of the Cauchy-Reimann equations. If we represented a complex function as $f(z) = u(r, \theta) + iv(r, \theta)$ where $z = re^{i\theta}$ then the following are equivalent to what we previously stated as the Cauchy-Reimann equations:

$$v_\theta = ru_r, \qquad u_\theta = -rv_r$$

When using polar coordinates, it can be shown that $f'(z) = e^{-i\theta}(u_r + iv_r)$.

The Cauchy-Riemann theorem and its converse hold true if we use the above form of the Cauchy-Reimann equations. Several proofs of this fact can be found at "Proof of Cauchy Riemann Equations in Polar Coordinates" [96].

For example, consider $f(z) = \frac{1}{z^n}$ for $n \in \mathbb{N}$. We first write $f(z)$ in polar form

$$\frac{1}{z^n} = \frac{1}{(re^{i\theta})^n} = \frac{1}{r^n}\frac{1}{e^{in\theta}} = \frac{1}{r^n}e^{-in\theta} = \frac{1}{r^n}(\cos(n\theta) - i\sin(n\theta))$$

So,

$$u(r, \theta) = \frac{\cos n\theta}{r^n}, \qquad v(r, \theta) = -\frac{\sin n\theta}{r^n}$$

Computing the Cauchy-Riemann equations, we see that both hold true, i.e.,

$$v_\theta = -\frac{n\cos n\theta}{r^n} = ru_r, \qquad u_\theta = -\frac{n\sin n\theta}{r^n} = -rv_r$$

So, $f'(z)$ exists at all $z \in \mathbb{C}$ and is given by

$$e^{-i\theta}(u_r + iv_r) = e^{-i\theta}\left(-\frac{n\cos n\theta}{r^{n+1}} + i\ \frac{n\sin n\theta}{r^{n+1}}\right) = -\frac{n}{r^{n+1}e^{i\theta}}(\cos n\theta - i\sin n\theta)$$

$$= -\frac{n}{r^{n+1}e^{i\theta}}\ e^{-n\theta} = -\frac{n}{(re^{i\theta})^{n+1}} = -\frac{n}{z^{n+1}}$$

4.5.6 Analytic Functions

A complex function $f(z)$ is said to be **analytic** (or holomorphic) at a point z_0 if it has a derivative at each point in some neighborhood of z_0 (where "neighborhood" is as defined for topological spaces in Section 3.5.2).

The function $f(z) = |z|^2$ (which we discussed previously) is not analytic at any point since its derivative only exists at $z = 0$ and not throughout any neighborhood of 0 or any other point in $\mathbb{C}$.

If a complex function $f(z)$ is not analytic at a given point z_0 but is analytic at some point in every neighborhood of z_0, then z_0 is referred to as a **singularity**, or singular point, of $f(z)$. For example, $1+i$ is a singularity of $f(z)=\frac{1}{z-(1+i)}$.

An **entire** complex function is one that is analytic for all $z \in \mathbb{C}$.

The following are some properties of analytic functions (stated without proof). Each statement is made under the assumption that the function(s) are analytic over a given set.

- The sums, products, and compositions of analytic functions are analytic. The quotient is analytic provided the function in the denominator does not vanish at any point.
    - The reciprocal of an analytic function that is nowhere zero is analytic. This is basically a subcase of the quotient property.
- If an analytic function f is invertible and has a derivative which is nowhere zero, then $f^{-1}(z)$ is analytic.
- Every analytic function is smooth, i.e., infinitely differentiable.
- (Liouville's theorem) An entire function that is bounded must be a constant.
- If $f(z)$ is analytic over a domain (i.e., an open connected set) D and if $f(z)=0$ everywhere in D, then $f(z)$ is constant in D.
- An analytic function over a domain (i.e., an open connected set) D is uniquely determined over D by its values in a subdomain S, or along a line segment, contained within D.

4.6 Elementary Functions

4.6.1 Polynomial and Rational Functions

Polynomials of a complex variable with complex coefficients are analytic everywhere in the complex plane. Complex **rational functions** (the quotient of two complex polynomials) are analytic everywhere except where the denominator is zero. For example, consider the following function:

$$f(z)=\frac{p(z)}{q(z)}=\frac{z^3+3\,z^2-13\,z-15}{z^4-11\,z^3+42\,z^2-68\,z+40}=\frac{(z+1)(z-3)(z+5)}{(z-5)(z-2)^3}$$

$f(z)$ is analytic everywhere except for $z=2$ and $z=5$ (the roots of the denominator). The degree of the polynomial in the numerator is 3, and the degree of the polynomial in the denominator is 4. The root $z=2$ of the denominator is known as a repeated root. The roots of the denominator are known as **poles**.

In the above expression for $f(z)$, we expressed the numerator and denominator in factored form. Such factorization is always possible for a complex polynomial due to the fundamental theorem of algebra (see Theorem 114). In particular, a complex polynomial $p_n(z)$ of degree n has a root (call it z_1).

Theorem 120. If a complex polynomial has real coefficients, then it can be expressed as the product of linear and quadratic factors, each having real coefficients.

Proof: Take $f(z) = z^n + a_1 z^{n-1} + a_2 z^{n-2} + \cdots + a_{n-1} z + a_n$ where the coefficients are real numbers, i.e., $a_i \in \mathbb{R}, i = 1,2, \ldots, n$.

From Theorem 113, we have that

$$\overline{f(z)} = \bar{z}^n + a_1 \bar{z}^{n-1} + a_2 \bar{z}^{n-2} + \cdots + a_{n-1} \bar{z} + a_n = f(\bar{z})$$

So, if $f(z_0) = 0$ then $f(\overline{z_0}) = \overline{f(z_0)} = 0$, i.e., complex roots of complex polynomials with real coefficients come in conjugate pairs.

From the fundamental theorem of algebra, we can write $f(z)$ in terms of its roots, i.e.,

$$f(z) = a(z - z_1)(z - z_2) \ldots (z - z_n)$$

If, for example, z_i and z_j are complex conjugates (i.e., $z_j = \overline{z_i}$) as well as roots of $f(z)$, we can replace $(z - z_i)\big(z - z_j\big) = (z - z_i)(z - \overline{z_\iota})$ with

$$z^2 - (z_i + \overline{z_\iota})z + z_i\, \overline{z_\iota} = z^2 - 2\big(Re(z_i)\big)z + |z_i|^2$$

So, we have replaced a pair of conjugate roots with a single quadratic with real coefficients. The other complex roots can be paired in a similar manner and reduced to quadratics with real coefficients. The real roots, if any, can be factored out as one-degree polynomial of the form $(z - z_k)$ where z_k is a real-valued root of $f(z)$. ■

For example, consider $f(z) = z^7 - 1$ which has real root 1, and complex roots

$$-\cos\left(\frac{\pi}{7}\right) \pm i \sin\left(\frac{\pi}{7}\right)$$

$$\sin\left(\frac{3\pi}{14}\right) \pm i \cos\left(\frac{3\pi}{14}\right)$$

$$\sin\left(\frac{\pi}{14}\right) \pm i \cos\left(\frac{\pi}{14}\right)$$

Using the formula for combining conjugate roots (developed in the proof of Theorem 120), we have

$$f(z) = (z - 1)\left(z^2 + 2\cos\left(\frac{\pi}{7}\right)z + 1\right)\left(z^2 - 2\sin\left(\frac{3\pi}{14}\right)z + 1\right)\left(z^2 - 2\sin\left(\frac{\pi}{14}\right)z + 1\right)$$

. . .

The polynomials that we have seen so far are all written in a form that is centered about $z = 0$. However, any polynomial can be rewritten to be centered around any point. For example, consider the polynomial $z^3 - z^2 - 4z + 9$. We can rewrite $f(z)$ to be centered around (for example) $z = 1$, with the form

$$f(z) = a_0 + a_1(z - 1) + a_2(z - 1)^2 + a_3(z - 1)^3$$

There are several ways to solve for the a_i terms. One way is to expand the alternate representation of $f(z)$, equate coefficients with the original representation and then solve for the a_i terms. An easier way is to notice the following:

$$a_0 = f(1) = 5$$

$$a_1 = f'(1) = -3$$

$$2!\, a_2 = f''(1) = 4 \;\Rightarrow\; a_2 = 2$$

$$3!\, a_3 = f'''(1) = 6 \Rightarrow a_3 = 1$$

So,

$$f(z) = (z-1)^3 + 2(z-1)^2 - 3(z-1) + 5$$

This is known as the Taylor form of $f(z)$ centered at $z = 1$.

In general, if $f(z)$ is complex polynomial of degree n, its Taylor form about $z = z_0$ is given by the expression

$$\sum_{i=0}^{n} \frac{f^i(z_0)}{i!}(z-z_0)^i = \frac{f(z_0)}{0!} + \frac{f'(z_0)}{1!}(z-z_0)^1 + \frac{f''(z_0)}{2!}(z-z_0)^2 + \cdots + \frac{f^n(z_0)}{n!}(z-z_0)^n$$

. . .

Recall the partial fractions technique from calculus for finding antiderivatives. The same technique can be used to simplify rational complex functions as a prelude to finding an antiderivative. For example, consider the following rational function:

$$f(z) = \frac{z^2 + 4z - 3}{(z+5)(z-3)^3}$$

In terms of partial fractions, we want to write $f(z)$ in the form

$$\frac{a}{z+5} + \frac{b_3}{(z-3)^3} + \frac{b_2}{(z-3)^2} + \frac{b_1}{z-3}$$

One approach is to determine a common denominator for the above expression, add the terms, equate coefficients with the numerator of the original expression, and then solve a system of four equations with four unknowns. The solution is

$$a = -\frac{1}{256}, \qquad b_3 = \frac{9}{4}, \qquad b_2 = \frac{31}{32}, \qquad b_3 = \frac{1}{256}$$

[**Author's Remark**: If you tried to verify the above result by hand, I'm sure you had a lot of fun. Well, not really – the problem is quite messy. I cheated and used Wolfram Alpha with the command line

expand into partial fractions (z^2+4z-3)/((z+5)(z-3)^3)

In any event, there is another approach that is a bit simpler.]

We can also find a as follows:

$$a = \lim_{z\to -5}(z+5)f(z) = \lim_{z\to -5}\frac{z^2+4z-3}{(z-3)^3} = \frac{25-20-3}{(-8)^3} = \frac{1}{256}$$

Similarly, we have that

$$b_3 = \lim_{z\to 3}(z-3)^3 f(z) = \lim_{z\to -5}\frac{z^2+4z-3}{z+5} = \frac{9+12-3}{8} = \frac{9}{4}$$

To find b_2, we need to take a derivative, i.e.,

$$\lim_{z\to 3}\frac{d}{dz}[(z-3)^3 f(z)] = \lim_{z\to 3}\frac{d}{dz}\left[\frac{a(z-3)^3}{z+5} + b_3 + b_2(z-3) + b_1(z-3)^2\right]$$

$$= \lim_{z\to 3}\frac{d}{dz}\left[\frac{a(z-3)^3}{z+5}\right] + \lim_{z\to 3}[b_2 + 2b_1(z-3)] = b_2$$

All the terms of $\frac{d}{dz}\left[\frac{a(z-3)^3}{z+5}\right]$ have a factor of $z-3$ and thus, it is 0 when evaluated at $z=3$.

Working from the original equation for $f(z)$, we also have that

$$\lim_{z\to 3}\frac{d}{dz}[(z-3)^3 f(z)] = \lim_{z\to 3}\frac{d}{dz}\left[\frac{z^2+4z-3}{z+5}\right] = \lim_{z\to 3}\frac{z^2+10z+23}{(z+5)^2} = \frac{31}{32}$$

Thus, $b_2 = \frac{31}{32}$.

To find b_1, we need to take a second derivative of $(z-3)^3 f(z)$ as $z\to 3$, i.e.,

$$\lim_{z\to 3}\frac{d^2}{dz^2}[(z-3)^3 f(z)] = \lim_{z\to 3}\frac{d^2}{dz^2}\left[\frac{a(z-3)^3}{z+5}\right] + 2b_1 = 2b_1$$

All the term of $\frac{d^2}{dz^2}\left[\frac{a(z-3)^3}{z+5}\right]$ have a factor of $z-3$ and thus, it is 0 when evaluated at $z=3$.

Working from the original equation for $f(z)$, we also have that

$$\lim_{z\to 3}\frac{d^2}{dz^2}[(z-3)^3 f(z)] = \lim_{z\to 3}\frac{d^2}{dz^2}\left[\frac{z^2+4z-3}{z+5}\right] = \lim_{z\to 3}\frac{4}{(z+5)^3} = \frac{1}{128}$$

Thus, $2b_1 = \frac{1}{128} \Rightarrow b_1 = \frac{1}{256}$.

. . .

In general, the following steps can be used to represent the rational function $f(z) = \frac{p(z)}{q(z)}$ in terms of partial fractions:

- Factor $q(z)$ in terms of its roots, i.e., $q(z) = (z - z_1)^{n_1}\,(z - z_2)^{n_2} \ldots (z - z_k)^{n_k}$ where z_i is a root of multiplicity n_i for $i = 1,2, \ldots, k$.
- Write the rational function in terms of partial fraction (with constants to be determined).

$$f(z) = \frac{a_{10}}{(z - z_1)^{n_1}} + \frac{a_{11}}{(z - z_1)^{n_1 - 1}} + \cdots + \frac{a_{1,n_1 - 1}}{z - z_1} +$$

$$\frac{a_{20}}{(z - z_2)^{n_2}} + \frac{a_{21}}{(z - z_2)^{n_2 - 1}} + \cdots + \frac{a_{2,n_2 - 1}}{z - z_2} +$$

$$\ldots$$

$$\frac{a_{k0}}{(z - z_k)^{n_k}} + \frac{a_{k1}}{(z - z_k)^{n_k - 1}} + \cdots + \frac{a_{k,n_k - 1}}{z - z_k}$$

- From here, we have two choices.
    a. We can determine a common denominator for the above expression, add the terms, equate the coefficients of the numerator of the resulting sum with the coefficients of $p(z)$, and then solve a system of equations for the various a_{ij} terms.
    b. Use the limit/derivative approach taken in the previous example. We have the following general formula, where $\frac{d^i}{dz^i}$ is the i^{th} derivative with respect to z and with $i = 0$ meaning "take no derivative".

$$a_{ji} = \lim_{z \to z_j} \frac{1}{i!} \frac{d^i}{dz^i} \left[\left(z - z_j\right)^{n_j} f(z) \right], \qquad i = 0,1, \ldots, n_j - 1, \qquad j = 1,2, \ldots, k$$

4.6.2 Exponential, Trigonometric and Hyperbolic Functions

The complex exponential function e^z is not a one-to-one mapping of the complex plane. [Recall that a function $f(z)$ is one-to-one if $f(z_1) = f(z_2)$ implies $z_1 = z_2$.] The reason is that e^z is a multi-valued function, as shown in the following theorem.

Theorem 121. (1) $e^z = 1$ if and only if $z = 2n\pi i$ for $n \in \mathbb{Z}$. (2) $e^{z_1} = e^{z_2}$ if and only if $z_1 = z_2 + 2n\pi i$ for $n \in \mathbb{Z}$.

Proof: [Recall $\mathbb{Z}$ is the set of all integers.]

We will use (1) to prove (2).

Concerning part (1) of the theorem, assume that $e^z = 1$. Letting $z = x + iy$, we have

$$|e^z| = \left|e^{x+iy}\right| = e^x = 1 \Rightarrow x = 0$$

So, $e^z = e^{iy} = \cos y + i \sin y = 1$ which implies $\cos y = 1$ and $\sin y = 0$, and this is only true when $y = 2n\pi$ for $n \in \mathbb{Z}$. Thus, $z = 2n\pi i$.

Conversely, if $z = 2n\pi i$ for $n \in \mathbb{Z}$, then $e^z = e^{2n\pi i} = \cos 2n\pi + i \sin 2n\pi = 1$.

For the proof of part (2), we first note that $e^{z_1} = e^{z_2}$ if and only if $e^{z_1 - z_2} = 1$. From part (1) of the theorem, $e^{z_1 - z_2} = 1$ if and only if $z_1 - z_2 = 2n\pi i$ for $n \in \mathbb{Z}$. ■

Based on the result of the above theorem, e^z is periodic with complex period $2\pi i$.

The complex exponential function is used to define the complex sine and cosine functions as follows:

$$\sin z = \frac{e^{iz} - e^{-iz}}{2i}, \qquad \cos z = \frac{e^{iz} + e^{-iz}}{2i}$$

Using the derivative of e^z (which we derived earlier), and a simple application of the chain rule, we can determine the derivative of $\sin z$.

$$\frac{d}{dz}\sin z = \frac{d}{dz}\left(\frac{e^{iz} - e^{-iz}}{2i}\right) = \frac{e^{iz} - (-1)e^{-iz}}{2i} = \cos z$$

Similarly, we have that $\frac{d}{dz}\cos z = -\sin z$.

Many of the trigonometric identities for real numbers hold true for complex trigonometric functions, e.g.,

$$\sin(z + 2\pi) = \frac{e^{i(z+2\pi)} - e^{-i(z+2\pi)}}{2i} = \frac{e^{iz}e^{2\pi i} - e^{-iz}e^{-2\pi i}}{2i} = \frac{e^{iz} - e^{-iz}}{2i} = \sin z$$

$$(\sin z)^2 + (\cos z)^2 = \left(\frac{e^{iz} - e^{-iz}}{2i}\right)^2 + \left(\frac{e^{iz} + e^{-iz}}{2}\right)^2$$

$$= \frac{-e^{2iz} + 2 - e^{-2iz} + e^{2iz} + 2 + e^{-2iz}}{4} = 1$$

. . .

The other complex trigonometric functions are defined in terms of $\sin z$ and $\cos z$, i.e.,

$$\tan z = \frac{\sin z}{\cos z}, \qquad \cot z = \frac{\cos z}{\sin z}, \qquad \sec z = \frac{1}{\cos z}, \qquad \csc z = \frac{1}{\sin z}$$

Derivatives of the above can be obtained using the quotient rule and the already established derivatives for $\sin z$ and $\cos z$. For example,

$$\frac{d}{dz}\tan z = \frac{\cos z \cos z + \sin z \sin z}{(\cos z)^2} = \frac{1}{(\cos z)^2} = (\sec z)^2$$

. . .

The complex hyperbolic functions are defined in an analogous manner to the real-valued hyperbolic functions, i.e.,

$$\sinh z = \frac{e^z - e^{-z}}{2}, \qquad \cosh z = \frac{e^z + e^{-z}}{2}$$

$$\tanh z = \frac{\sinh z}{\cosh z}, \qquad \coth z = \frac{\cosh z}{\sinh z}, \qquad \operatorname{sech} z = \frac{1}{\cosh z}, \qquad \operatorname{csch} z = \frac{1}{\sinh z}$$

In addition, we have the following identities:

$$\sin(iz) = i\sinh z\,, \qquad \sinh(iz) = i\sin z\,, \qquad \cos(iz) = \cosh z\,, \cosh(iz) = \cos z$$

$$(\cosh z)^2 - (\sinh z)^2 = 1$$

4.6.3 Logarithmic Function

The complex logarithmic function is defined as the inverse of the complex exponential function. If $z = e^w$ where $w = u + iv$, then we define $w = \log z$. Since e^w is never zero, we have that $z \neq 0$ as a condition on $\log z$. If we write z in polar form, then we have

$$z = re^{i\theta} = e^{u+iv} = e^u e^{iv}$$

Thus, $r = e^u$ which implies $u = \ln r$ where ln is the real-valued natural logarithm.

Further, $v = \arg z = \theta$, but arg is a multi-valued parameter (see the definition in Section 4.2). So,

$$v = Arg(z) + 2\pi n, \qquad n \in \mathbb{Z}$$

Based on the above analysis, we define the complex logarithmic function as follows:

For $z \neq 0$, the **complex logarithm** is a multi-valued function defined by

$$\log z = \ln|z| + iArg(z) + i2\pi n, \qquad n \in \mathbb{Z}$$

Since, by definition, the principal branch of $\arg(z)$ has a branch cut along the negative real axis, $\log z$ also has a branch cut (ray of discontinuity) along the negative real axis.

The **principal value** of $\log z$ is taken when $n = 0$ in the above equation. Further, $\operatorname{Log} z = \ln|z| + i\,Arg(z)$ is defined to be the **principal branch** of $\log z$.

For example,

$$\log(1 + i\sqrt{3}) = \ln(2) + \frac{i\pi}{3} + + i2\pi n, \qquad n \in \mathbb{Z}$$

Note that the polar form of $1 + i\sqrt{3}$ is

$$2e^{\frac{i\pi}{3}}$$

The following properties hold true for the complex logarithmic function:

$$\log(z_1 z_2) = \log z_1 + \log z_2$$

$$\log\left(\frac{z_1}{z_2}\right) = \log z_1 - \log z_2$$

Because the logarithmic function has multiple branches, the above formulas can be a bit tricky to apply. For example, consider $z_1 = z_2 = -1 - i$. On the one hand, we have

$$\log(-1-i) = \ln(-1-i) + \frac{5\pi}{4} + i2n\pi = \ln\sqrt{2} + \frac{5\pi}{4} + i2n\pi, \qquad n \in \mathbb{Z}$$

and so,

$$\log(-1-i) + \log(-1-i) = 2\ln\sqrt{2} + \frac{5\pi}{2} + i2m\pi = \ln 2 + \frac{5\pi}{2} + i2m\pi, \qquad m \in \mathbb{Z}$$

On the other hand,

$$\log\big((-1-i)(-1-i)\big) = \log(2i) = \ln 2 + \frac{\pi}{2} + i2s\pi, \qquad s \in \mathbb{Z}$$

The two results don't match unless we make the proper selection of the branch of the log function in each case. There are multiple correct solutions, e.g., take $m = -1$ and $s = 0$.

. . .

Since the definition of $\log z$ entails the use of $\arg z$, and $\arg z$ is discontinuous along the negative real axis (recall the discussion surrounding Figure 60), we concluded that $\log z$ is discontinuous along the negative real axis. Further, since $\log z$ is undefined for $z = 0$, that gives us one additional point of discontinuity.

Similar to the real function, the derivative of $\log z$ is $\frac{1}{z}$ for all z except 0 and the negative real axis. The proof goes as follows:

If we let $\log z = w$, then $e^w = z$ or $e^{\log z} = z$. Taking the derivative of both sides with respect to z, we have

$$\frac{d}{dz} e^{\log z} = \frac{d}{dz} z = 1$$

Applying the chain rule to the left side of the above equation, we get

$$e^{\log z} \cdot \frac{d}{dz} \log z = 1 \Rightarrow z \cdot \frac{d}{dz} \log z = 1 \Rightarrow \frac{d}{dz} \log z = \frac{1}{z}$$

4.7 Complex Integration

4.7.1 Contours

For real valued functions, one typically defines a definite integral over an interval, e.g.,

$$\int_{-1}^{3} x^3\,dx = \frac{x^4}{4}\Big]_{-1}^{3} = \frac{81}{4} - \frac{1}{4} = \frac{80}{4} = 20$$

Just for the record, a real-valued function does not need to be continuous for the Riemann integral to exist. From the Wikipedia article on the Riemann integral [97]:

> A bounded function on a compact interval $[a, b]$ is Riemann integrable if and only if it is continuous almost everywhere (the set of its points of discontinuity has measure zero, in the sense of Lebesgue measure). This is the Lebesgue-Vitali theorem (of characterization of the Riemann integrable functions). It has been proven independently by Giuseppe Vitali and by Henri Lebesgue in 1907, and uses the notion of measure zero, but makes use of neither Lebesgue's general measure or integral.

While it is possible to define definite integrals for complex functions, the situation is more complex (no pun intended). If one is to define a definite integral of complex function $f(z)$ from z_1 to z_2, it is necessary to make clear which of the infinite number of paths one wants to follow in going from z_1 to z_2. In most cases, it is sufficient to select either a line segment or a circular arc as the path between z_1 to z_2. However, a more general approach is taken when stating theorems concerning the integration of complex functions. In particular, the path from z_1 to z_2 is required to be a contour where a contour is a finite sequence of directed smooth curves (which we will define more precisely below).

We start with the definitions of “smooth arc” and “smooth closed curve”, both of which are classified as smooth curves.

A set S in the complex plane is a **smooth arc** if there exists a function $p(t)$ that maps the real interval $[a, b]$ to S such that

i. $p(t)$ has a continuous derivative on $[a, b]$

ii. $p(t) \neq 0\ \forall t \in [a, b]$

iii. $p(t)$ is one-to-one on $[a, b]$

A set S is a **smooth closed curve** if it is the codomain of some continuous complex-valued function $p(t), a \leq\ t \leq b$ satisfying conditions (i) and (ii) above and such that $p(t)$ is one-to-one on interval $[a, b)$, but $p(b) = p(a)$ and $p'(a) = p'(b)$.

Figure 65 depicts a simple example of a smooth arc that maps the real interval $[0,1]$ to the line segment between $-2-2i$ and $2+2i$.

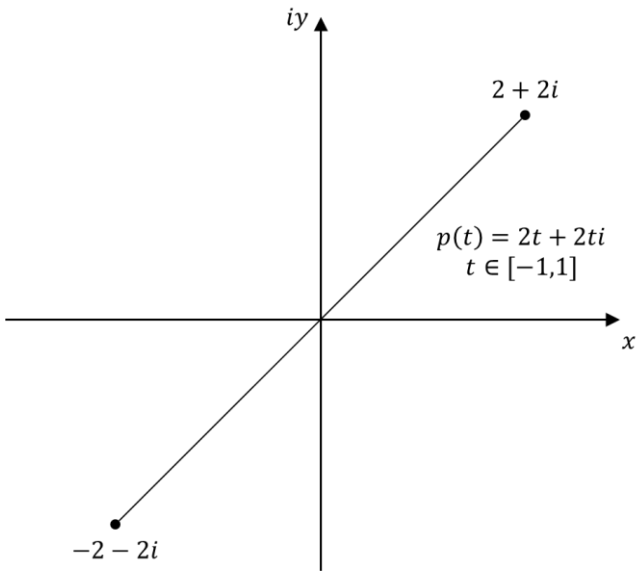


Figure 65. Smooth arc – straight line example

An example of a semicircle smooth arc is shown Figure 66. Two different parameterizations are provided (one using polar coordinates and the other using Cartesian coordinates). For a closed smooth arc example, just take $p_1(\theta)$ from our semicircle example and extend the parameter range to $0 \le \theta \le 2\pi$.

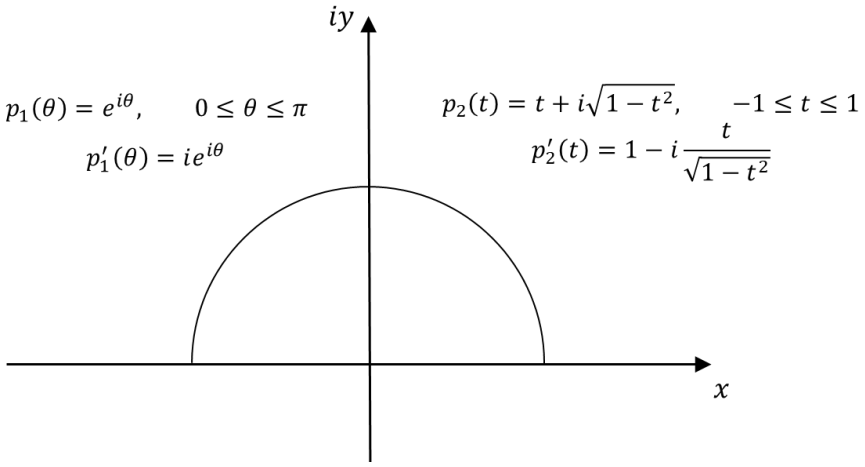


Figure 66. Smooth arc – semicircle example

The parametrization of a curve provides a natural ordering of points on the curve. This suggests the notion of a **directed smooth curve**. For example, the parameterization in Figure 65 implies that the curve starts at the point $-2-2i$ and ends at the point $2+2i$. If we wanted to go in the opposite direction, we use the parameterization $q(t) = p(-t) = -2t - 2ti, -1 \le t \le 1$. For smooth arcs, there are only two possible directions. In general, if $p(t),\ a \le t \le b$, is a valid parametrization of

a directed smooth curve, then $p(-t), -b \leq t \leq -a$ is a valid parameterization going in the opposite direction.

However, for smooth closed curves, we can start at any point on the curve and then go in either of two possible directions. Thus, there are an infinite number of valid parameterizations in this case.

. . .

A **contour** Γ is either a single point or a finite sequence of directed smooth curves $S_1, S_2, \dots, S_n$ such that the endpoint of S_k coincides with the initial point of S_{k+1} for $k = 1,2, \dots, n-1$. In terms of notation, we have $\Gamma = S_1 + S_2 + \cdots + S_n$. A **closed contour** is one in which its initial and terminal points coincide. Further, a **simple closed contour** (or loop) is a closed contour with no overlapping points other than its initial and terminal points.

While it may seem obvious, a simple closed contour (loop) always divides the plane into two regions, i.e., the interior and exterior of the loop. This fact is known as the **Jordan curve theorem** which we state below.

Theorem 122. Every simple closed contour (loop) separates the plane into two domains (open connected sets), each having the curve as its boundary. One of the domains is bounded (known as the interior of the loop) and the other domain is unbounded (known as the exterior of the loop).

The Jordan curve theorem is easy to state but exceedingly difficult to prove. From the Wikipedia article on this topic [94]:

> The statement of the Jordan curve theorem may seem obvious at first, but it is a rather difficult theorem to prove. Bernard Bolzano was the first to formulate a precise conjecture, observing that it was not a self-evident statement, but that it required a proof. It is easy to establish this result for polygons, but the problem came in generalizing it to all kinds of badly behaved curves, which include nowhere differentiable curves, such as the Koch snowflake and other fractal curves, or even a Jordan curve of positive area constructed by Osgood (1903).
>
> The first proof of this theorem was given by Camille Jordan in his lectures on real analysis, and was published in his book Cours d'analyse de l'École Polytechnique (1887).

Consider walking along a simple closed contour (loop) drawn on the ground. When the interior domain lies to your left, the loop is said to be **positively oriented**; otherwise, the loop is said to be oriented negatively. A positive orientation generalizes the concept of counterclockwise.

4.7.2 Contour Integrals

As we saw, the two-dimensional nature of the complex plane required a broader concept of a "limit" and also of "derivative" since the variable can approach its limit from an infinite number of directions. This two-dimensional characteristic also impacts the theory of integration, requiring us to consider integrals along general curves in the plane (i.e., the contours defined in the previous section), rather than just segments of the x-axis. Fortunately, familiar techniques from basic calculus such as using antiderivatives to evaluate integrals still apply in the case of complex functions.

Recall from the calculus of real functions that definite integrals were developed using Riemann sums. A brief review is as follows:

> Let $f: [a, b] \to \mathbb{R}$ be a function defined on a closed interval $[a, b] \in \mathbb{R}$, and let $\mathcal{P} = (x_0, x_1, \dots, x_n)$ be a partition of $[a, b]$ such that
>
> $$a = x_0 < x_1 < x_2 < \cdots < x_n = b$$
>
> A Riemann sum $\sigma_{\mathcal{P}}$ over $[a, b]$ with partition $\mathcal{P}$ is defined as
>
> $$\sigma_{\mathcal{P}} = \sum_{i=1}^{n} f(x_i^*)\Delta x_i\,, \qquad \Delta x_i = x_i - x_{i-1}, \qquad x_i^* \in (x_{i-1}, x_i)$$
>
> The Riemann sum may differ depending on the choice of the x_i^* terms. However, as $\Delta x_i \to 0$ for $\mathcal{P}$, $\sigma_{\mathcal{P}}$ will approach a specific value if the definite integral exists.

With some modifications, we can define a Riemann sum for a complex function over a directed smooth curve. For $n \in \mathbb{N}$, define a partition $\mathcal{P}_n$ of a directed smooth curve γ to be a set of points $z_0 = \alpha, z_1, z_2, \dots, z_n = \beta$ on γ where α is the initial point of γ and β is the terminal point of γ such that the following holds true:

- z_{i-1} precedes z_i as one traverses $\mathcal{P}_n$.
- Let $\mu(\mathcal{P}_n)$ equals the maximum $\Delta z_i = |z_i - z_{i-1}|$.
- z_i^* is any point on γ between z_{i-1} and z_i.

The situation is shown in Figure 67.

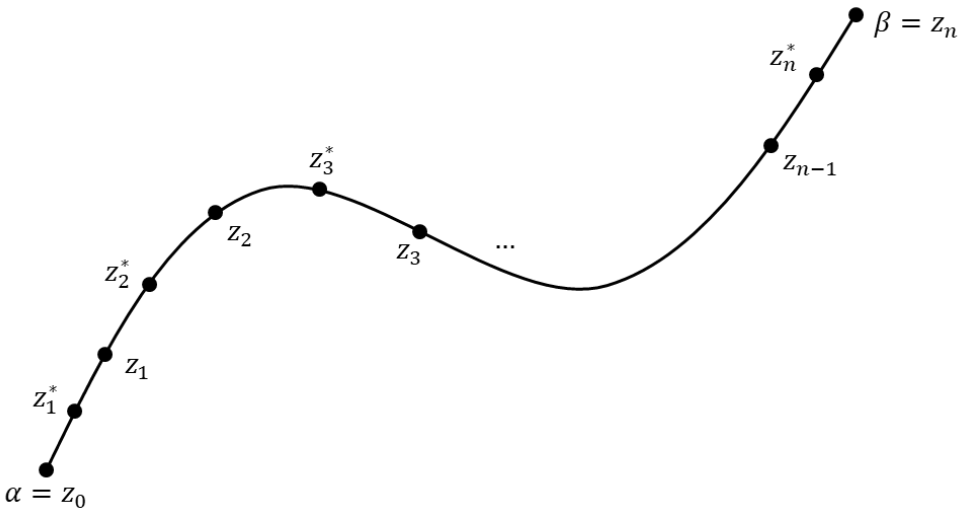


Figure 67. Partition of a smooth curve

For a function $f(z)$ defined on a directed smooth curve γ, the Riemann sum with respect to the partition $\mathcal{P}_n$ of γ is

$$\sigma(\mathcal{P}_n) = \sum_{i=1}^{n} f(z_i^*)\Delta z_i$$

$f(z)$ is said to be integrable along γ if there exists a complex number L such that the limit of every sequence of Riemann sums $\sigma(\mathcal{P}_1), \sigma(\mathcal{P}_2), \ldots, \sigma(\mathcal{P}_n), \ldots$ corresponding to any sequence of partitions of γ such that $\lim_{n\to\infty} \mu(\mathcal{P}_n) = 0$, i.e.,

$$\lim_{n\to\infty} \sigma(\mathcal{P}_n) = L \text{ whenever } \lim_{n\to\infty} \mu(\mathcal{P}_n) = 0$$

In this case, we say that L is the value of the integral of $f(z)$ along γ. This is represented as

$$L = \int_{\gamma} f(z)\, dz$$

The following basic properties hold true for the integral of a function over a directed smooth curve:

$$\int_{\gamma} [f(z) \pm g(z)]\, dz = \int_{\gamma} f(z)\, dz \pm \int_{\gamma} g(z)\, dz$$

$$\int_{\gamma} cf(z)\, dz = c \int_{\gamma} f(z)\, dz\,, c \in \mathbb{C}$$

$$\int_{-\gamma} f(z)\, dz = -\int_{\gamma} f(z)\, dz$$

In the last equation above, $-\gamma$ is the same as γ but directed in the opposite direction.

In the case of a function $f(z)$ whose domain is a smooth curve γ, we have a modified definition of continuity, i.e.,

> A function $f(z)$ having smooth curve γ as its domain of definition is continuous on γ if for any point $z_0 \in \gamma$ and for every $\epsilon > 0$, there exists $\delta > 0$ such that $|f(z) - f(z_0)| < \epsilon$ whenever $z \in \gamma$ and $|z - z_0| < \delta$.

With the above definition of continuity in hand, we have the following theorem.

Theorem 123. If f is continuous on the directed smooth curve γ, then f is integrable along γ.

The following theorem provides a method for evaluating definite integrals of a complex function over a directed smooth curve.

Theorem 124. If f is a continuous function on a directed smooth curve γ and $p(t), a \le t \le b$ is any valid parametrization of γ consistent with its direction, then

$$\int_{\gamma} f(z)\, dz = \int_{a}^{b} f(p(t))\, p'(t)\, dt$$

. . .

As an example, we will compute the value of $\int_{\gamma} \bar{z}\, dz$ where γ is the right half of the circle $|z| = 2$, i.e.,

$$p(\theta) = 2e^{i\theta}, -\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}$$

Applying Theorem 124, we have

$$L = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \overline{p(\theta)}\, p'(\theta) d\theta = 4\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} e^{-i\theta} i e^{i\theta}\, d\theta = 4i\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} d\theta = 4\pi i$$

. . .

For our next example, we will integrate the function $f(z) = (z - z_0)^n$ over the circle γ_r given by $p(\theta) = z_0 + re^{i\theta}, 0 \leq \theta \leq 2\pi$. We first compute the input to the formula in Theorem 124:

$$f\big(p(\theta)\big) = \big(z_0 + re^{i\theta} - z_0\big)^n = r^n e^{in\theta}$$

$$p'(\theta) = ire^{i\theta}$$

Inserting the above into the integral formula, we have

$$\int_{\gamma_r} (z - z_0)^n\, dz = \int_0^{2\pi} \big(r^n e^{in\theta}\big)\big(ire^{i\theta}\big)\, d\theta = ir^{n+1}\int_0^{2\pi} e^{i(n+1)\theta}\, d\theta$$

If $n \neq -1$, then the above definite integral evaluates as follows:

$$ir^{n+1}\frac{e^{i(n+1)\theta}}{i(n+1)}\Bigg]_0^{2\pi} = ir^{n+1}\left[\frac{1}{i(n+1)} - \frac{1}{i(n+1)}\right] = 0$$

If $n = -1$, then we have

$$ir^{n+1}\int_0^{2\pi} e^{i(n+1)\theta}\, d\theta = i\int_0^{2\pi} d\theta = 2\pi i$$

. . .

Consider $f(z) = z$ defined over any direct smooth curve γ between points z_1 and z_2. Let $p(t), a \leq t \leq b$ be a valid parameterization of γ so that $p(a) = z_1$ and $p(b) = z_2$. From the formula in Theorem 124, we have that

$$L = \int_{\gamma} f(z)\, dz = \int_a^b f(p(t))\, p'(t)\, dt = \int_a^b p(t)\, p'(t)\, dt$$

Since $\frac{d}{dz}\frac{[p(t)]^2}{2} = p(t)p'(t)$ (by the chain rule), we know the antiderivative of the above integral, i.e.,

$$L = \frac{[p(t)]^2}{2}\Big]_a^b = \frac{p(b)^2 - p(a)^2}{2} = \frac{z_2^2 - z_1^2}{2}$$

So, the integral only depends on the endpoints of γ and is independent of the path taken between the endpoints.

. . .

The contour integral of a continuous function f over a contour $\Gamma = \gamma_1 + \gamma_2 + \cdots + \gamma_n$ can be defined in terms of the directed smooth curves comprising the contour, i.e.,

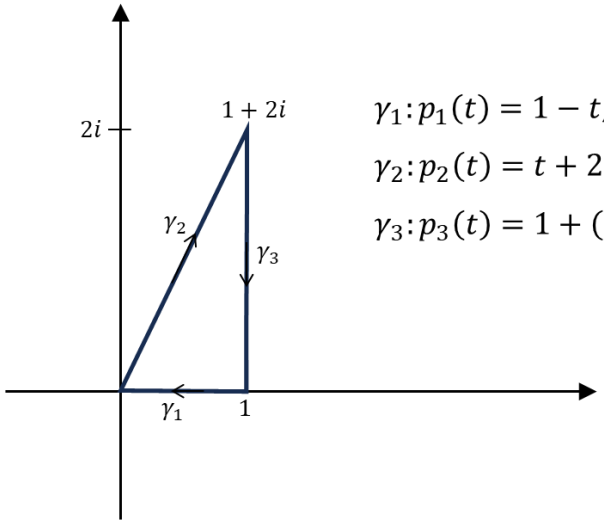
$$\int_{\Gamma} f(z)\,dz = \int_{\gamma_1} f(z)\,dz + \int_{\gamma_2} f(z)\,dz + \cdots + \int_{\gamma_n} f(z)\,dz$$

In the case that the contour Γ_0 consists of a single point, we define the integral to be zero, i.e.,

$$\int_{\Gamma_0} f(z)\,dz = 0$$

. . .

Consider the function $f(z) = \bar{z}^2$ over contour Γ with the parameterization shown in



$$\gamma_1 : p_1(t) = 1 - t, 0 \le t \le 1$$
$$\gamma_2 : p_2(t) = t + 2ti, 0 \le \mathrm{t} \le 1$$
$$\gamma_3 : p_3(t) = 1 + (2 - t)i, 0 \le t \le 2$$

Figure 68. Triangular shaped contour

The integral of $f(z)$ over Γ can be determined by evaluating the integral over each of the three directed smooth curves that comprise Γ, i.e.,

$$\int_{\Gamma} \bar{z}^2\,dz = \int_{\gamma_1} \bar{z}^2\,dz + \int_{\gamma_2} \bar{z}^2\,dz + \int_{\gamma_3} \bar{z}^2\,dz$$

$$\int_{\gamma_1} \bar{z}^2\, dz = \int_0^1 \overline{p_1(t)}^2\, p_1'(t)\, dt = \int_0^1 (1-t)^2(-1)\, dt = \int_0^1 -t^2 + 2t - 1\, dt$$

$$= \left[-\frac{t^3}{3} + t^2 - t\right]_0^1 = -\frac{1}{3} + 1 - 1 = -\frac{1}{3}$$

$$\int_{\gamma_2} \bar{z}^2\, dz = \int_0^1 \overline{p_2(t)}^2\, p_2'(t)\, dt = \int_0^1 (t-2ti)^2(1+2i)\, dt = (1+2i)(1-2i)^2 \int_0^1 t^2\, dt$$

$$= 5(1-2i)\left[\frac{t^3}{3}\right]_0^1 = 5 \cdot \frac{1}{3}(1-2i) = \frac{5}{3} - \frac{10}{3}i$$

$$\int_{\gamma_3} \bar{z}^2\, dz = \int_0^2 \overline{p_3(t)}^2\, p_3'(t)\, dt = \int_0^2 (1-(2-t)i)^2(-i)\, dt = -i\left(-\frac{2}{3} - 4i\right) = -4 + \frac{2}{3}i$$

Putting the above results together, we have

$$\int_{\Gamma} \bar{z}^2\, dz = -\frac{1}{3} + \frac{5}{3} - \frac{10}{3}i - 4 + \frac{2}{3}i = -\frac{8}{3}(1+i)$$

4.7.3 Independence of Path

In some cases, it is possible to evaluate a contour integral by simply taking the difference of the antiderivative evaluated at the two endpoints of the contour, as stated in the following theorem (which is basically the fundamental theorem of calculus extended to contour integrals).

Theorem 125. If f(z) is continuous in a domain (i.e., open connected set) D and has antiderivative $F(z)$ for all $z \in D$, then for any contour $\Gamma \subset D$, with initial point z_1 and terminal point z_2, the following holds true

$$\int_{\Gamma} f(z)\, dz = F(z_2) - F(z_1)$$

Proof: See Theorem 6 in Section 4.3 of "Fundamentals of Complex Analysis" [95]. ■

The above theorem does not apply to our example from the previous section since $f(z) = \bar{z}^2$ is not analytic. (The function $f(z) = \bar{z}^2$ is not analytic since the Cauchy-Reimann equations do not hold true). Further, recall from Section 4.5.6 that analytic functions are infinitely differentiable. So, if $f(z)$ had an antiderivative $F(z)$, $F(z)$ and all of its derivatives would be analytic (a contradiction to $f(z)$ not being analytic).

If Γ is a simple closed contour (loop) within the domain of definition of a continuous function f which has antiderivative F for all of D, then it follows by Theorem 125 that $\int_{\Gamma} f(z)\, dz = 0$.

The following theorem relates several concepts that we have discussed so far.

Theorem 126. For a continuous function f in a domain (open connected set) D, the following conditions are equivalent:

- f has an antiderivative in D.
- $\int_{\Gamma} f(z)\,dz = 0$ for every loop (simple closed contour) Γ in D.
- For any two contours Γ_1 and Γ_2 in D that share the same initial and terminal points, $\int_{\Gamma_1} f(z)\,dz = \int_{\Gamma_2} f(z)\,dz$.

Proof: See Theorem 7 in Section 4.3 of "Fundamentals of Complex Analysis" [95]. ■

. . .

As an example of where Theorem 125, consider $f(z) = \sin z$ over the contour shown in Figure 69. The antiderivative of f is $F(z) = -\cos z$ and so, the integral over contour Γ is given by

$$\int_{\Gamma} \sin z \; dz = -\cos(2) + \cos(-2)$$

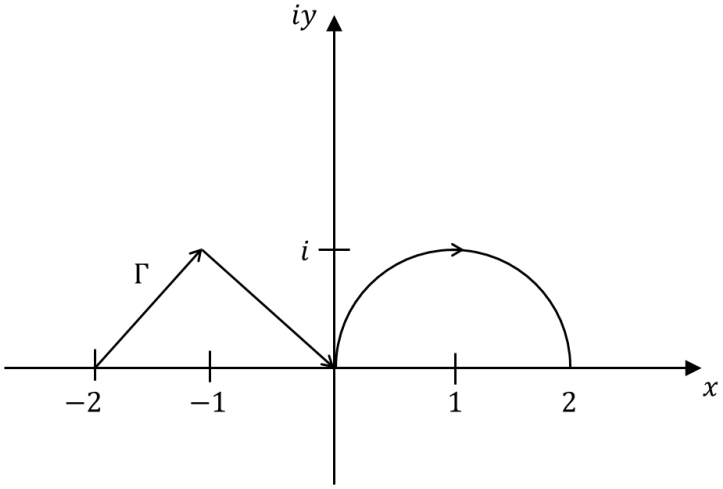


Figure 69. Contour example

. . .

Consider the function $f(z) = 1/z$ and the contour Γ shown in Figure 70.

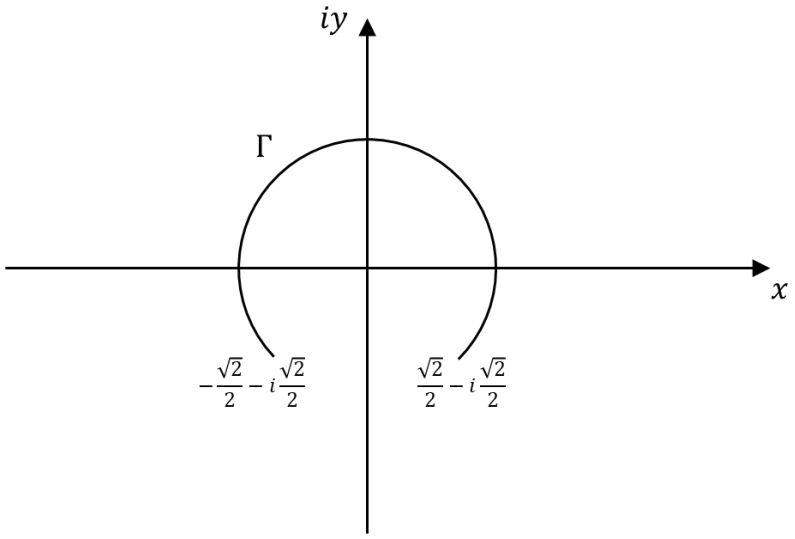


Figure 70. Contour based on portion of unit circle with center at the origin

The antiderivative of $\frac{1}{z}$ is $\log z$ but we have an issue with the definition of $\log z$, i.e., there is a branch cut along the negative real axis (which intersects the contour Γ). In order to use Theorem 125, we define the branch cut for $\log z$ to be the negative imaginary axis. In this case, we have

$$\operatorname{Log} z = \ln|z| + i\, Arg(z), \qquad \frac{3\pi}{2} < Arg(z) \leq -\frac{\pi}{2}$$

We can now apply the theorem to get

$$\int_{\Gamma} \frac{1}{z}\, dz = \operatorname{Log}\left(-\frac{\sqrt{2}}{2} - i\frac{\sqrt{2}}{2}\right) - \operatorname{Log}\left(\frac{\sqrt{2}}{2} - i\frac{\sqrt{2}}{2}\right) = \frac{5\pi}{4}i - \left(-\frac{\pi}{4}\right)i = \frac{3\pi}{2}i$$

In the above evaluation of the integral, we made use of the fact that

$$\ln\left|-\frac{\sqrt{2}}{2} - i\frac{\sqrt{2}}{2}\right| - \ln\left|\frac{\sqrt{2}}{2} - i\frac{\sqrt{2}}{2}\right| = \ln 1 - \ln 1 = 0 - 0 = 0$$

4.7.4 Cauchy's Integral Theorem and Formula

We start this section with a theorem concerning the invariance of an integral over different loops (i.e. simple closed contours) that are topologically equivalent in the sense that one can be continuously deformed into the other.

Theorem 127. (Deformation Invariance Theorem) If f is an analytic function in domain D containing loops Γ_0 and Γ_1 such that the loops can be continuous deformed into one another then

$$\int_{\Gamma_0} f(z)dz = \int_{\Gamma_1} f(z)dz$$

Proof: See Theorem 8 in Section 4.4 of "Fundamentals of Complex Analysis" [95] for a sketch of a proof. ■

Figure 71 depicts several loops that can be deformed into one other, assuming the region of interest is the entire complex plane. For example, we can contract Γ_1 to a point and then move it to the position of Γ_0. We can then expand the point Γ_0 to a circle of the same dimensions and direction as Γ_2. The method for changing direction is to shrink to a point and then expand with the direction reversed.

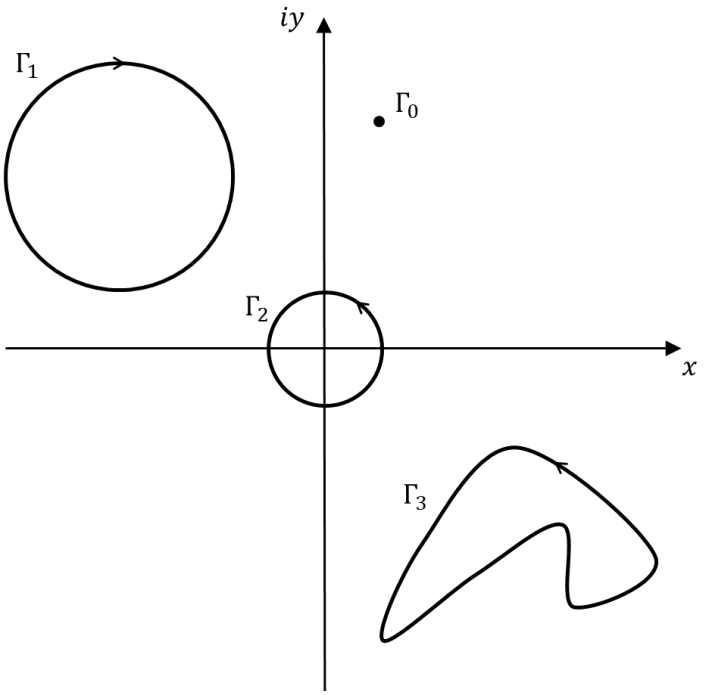


Figure 71. Equivalent loops under continuous deformation

Figure 72 depicts a domain D (gray area) with two circular areas removed. In this configuration, there is no way to continuously deform any of the loops in set $\{\Gamma_0, \Gamma_1, \Gamma_2\}$ into one of the other loops in the same set while remaining within the domain D. However, it is possible to deform loop Γ_0 into loop Γ_4 and vice versa.

A domain (open connected set) having the property that every loop can be continuously deformed to a point is called a **simply connected domain**. This is a special case of the definition of simply connected in Section 3.5.8.

Domain D in Figure 72 is not simply connected since we cannot deform contour Γ_1 or Γ_2 to a point. The interior of a disc, e.g., $|z - i| < 3$, is an example of a simply connected domain.

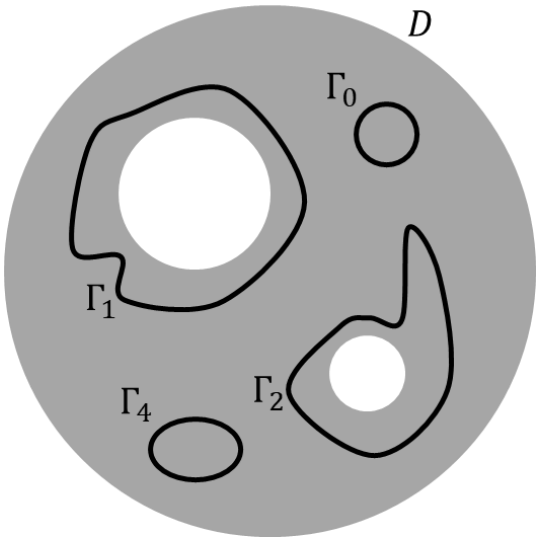


Figure 72. Domain that is not simply connected

Cauchy's integral theorem follows directly from the deformation invariance theorem.

Theorem 128. (Cauchy's Integral Theorem) If f is an analytic function in a simply connected domain D and Γ is any loop (closed contour) in D, then $\oint_{\Gamma} f(z)\, dz = 0$.

Proof: In a simply connected domain, any loop can be deformed into a single point, and we previously defined the integral of a function at a point to be zero. ■

The following theorem summarizes several of the previous results.

Theorem 129. An analytic function f, in a simply connected domain D, has an antiderivative, has contour integrals that are independent of path, and all its loop integrals are equal to zero.

Proof: That the loop integrals of f all equal zero follows from Theorem 128.

The path independence of contour integral for f follows from Theorem 127.

Since the loop integrals of f all vanish (equal zero), we have from Theorem 126 that f has an antiderivative in all of D. ■

. . .

As an example, consider the integral of the function $f(z) = \frac{e^z}{z^2-16}$ over the contour $|z| = 2$. In the domain $|z| < 3$, $f(z)$ is continuous and has derivative

$$f'(z) = \frac{e^z(z^2-16) - e^z(2z)}{(z^2-16)^2} = e^z \frac{z^2-2z-16}{(z^2-16)^2}$$

Since $f(z)$ is analytic in the simply connected domain $|z| < 3$, we have by Theorem 129

$$\int_{|z|=2} \frac{e^z}{z^2-16}\, dz = 0$$

. . .

The next example makes use of several of the preceding results. Our task is to find the integral of

$$f(z) = \frac{2z-5}{z^2-2z+2}$$

over the contour Γ shown in Figure 73. Contour Γ is divided into two parts, i.e., Γ_1 above the x axis, and Γ_2 below the x axis. The exact specification of the contour is not needed to determine the integral.

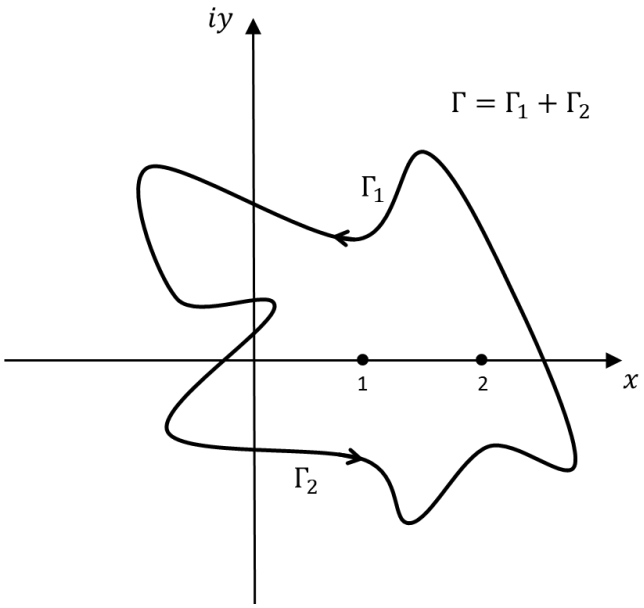
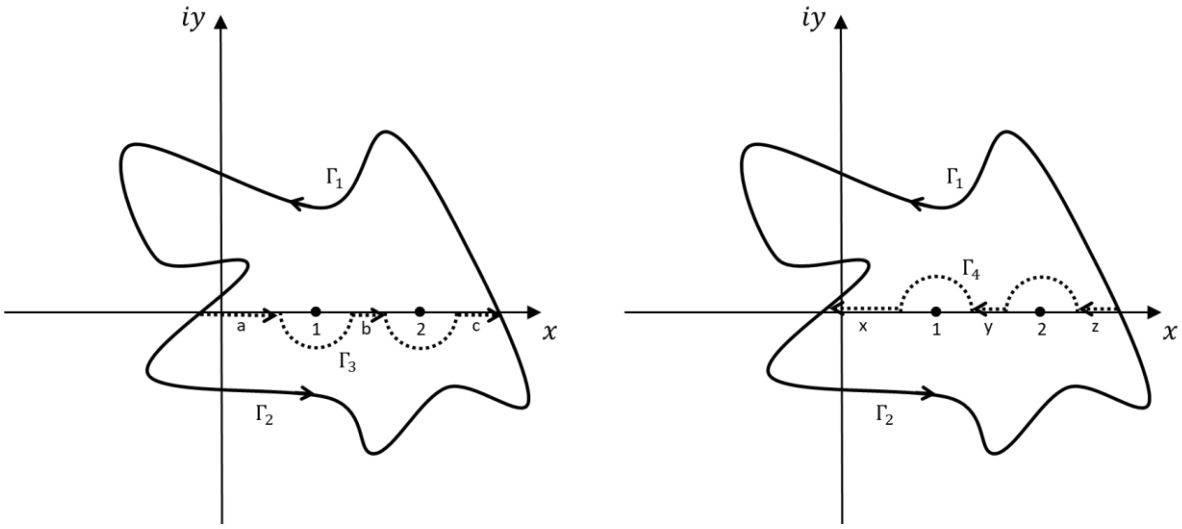


Figure 73. Contour about singularities

By the deformation invariance theorem, the value of the integral of f over Γ_2 is the same as the integral over Γ_3 (see the left side of the figure below). Similarly, the value of the integral of f over Γ_1 is the same as the integral over Γ_4 (see the right side of the figure below). Thus,

$$\int_{\Gamma} f(z)dz = \int_{\Gamma_3} f(z)dz + \int_{\Gamma_4} f(z)dz$$



There are several cancellations in the integration, i.e., the integral over a cancels the integral over x, the integral over b cancels the integral over y and the integral over c cancels the integral over z.

Assume the parts of the contours around points 1 and 2 are semicircles that form two circular contours, i.e., C_1 about point 1 and C_2 about point 2. Thus, our task is reduced to evaluating

$$\int_{C_1} f(z)dz + \int_{C_2} f(z)dz$$

Next, we represent $f(z)$ in terms of partial fractions.

$$\frac{2z-5}{z^2-2z+2} = \frac{A}{z-1} + \frac{B}{z-2}$$

Using the technique described earlier for the determination of the coefficients in a partial fraction expansion, we have

$$A = \lim_{z\to 1}(z-1)f(z) = 3$$

$$B = \lim_{z\to 2}(z-2)f(z) = -1$$

Previously, we computed the following integral where γ_r is a circle centered about z_0.

$$\int_{\gamma_1} \frac{1}{z-z_0}\, dz = 2\pi i$$

Thus, we have

$$\int_{\Gamma} f(z)dz = \int_{C_1} f(z)dz + \int_{C_2} f(z)dz$$

$$= \int_{C_1} \frac{3}{z-1} - \frac{1}{z-2}\,dz + \int_{C_2} \frac{3}{z-1} - \frac{1}{z-2}\,dz = 3(2\pi i) + 0 + 0 - (2\pi i) = 5\pi i$$

. . .

Related to Cauchy's integral theorem is something known as **Cauchy's integral formula**, which we state in the following theorem.

Theorem 130. If a function $f(z)$ is analytic in a simply connected domain D containing the positively oriented simple closed contour (loop) Γ and w is any point inside of Γ, then

$$f(w) = \frac{1}{2\pi i}\int_{\Gamma} \frac{f(z)}{z-w}\,dz$$

Proof: See Theorem 14 in Section 4.5 of "Fundamentals of Complex Analysis" [95]. ■

We can use Cauchy's integral formula to compute the following integral over the contour Γ defined by the circle $|z-(i+1)| = 3$ traversed once in the positive (counterclockwise) direction.

$$\int_{\Gamma} \frac{e^z + z^2 - 2z}{z-i}\,dz$$

$f(z) = e^z + z^2 - 2z$ is analytic over the entire complex plane and thus, analytic on and inside of Γ. Applying Cauchy's integral formula and noting the i is inside Γ, we have

$$\int_{\Gamma} \frac{e^z + z^2 - 2z}{z-i}\,dz = 2\pi i\, f(i) = 2\pi i\left(e^i - 1 - 2i\right)$$

. . .

By taking repeated derivatives of Cauchy's integral formula, we get what is known as the generalized Cauchy integral formula.

Theorem 131. If function f is analytic on and inside a positively oriented simple closed contour Γ and w is any point inside of Γ, then

$$\boldsymbol{f^{(n)}(w) = \frac{n!}{2\pi i}\int_{\Gamma} \frac{f(z)}{(z-w)^{n+1}}\,dz\,,\, n \in \mathbb{N}}$$

Further, if is analytic in a domain D, then all its derivatives exist and are analytic in D.

(The notation $f^{(n)}(w)$ means the n^{th} derivative of the function f with respect to w.)

We can use generalized Cauchy integral formula to evaluate the following modification of the previous example, with Γ as defined previously.

$$\int_{\Gamma} \frac{e^z + z^2 - 2z}{(z-i)^3}\,dz$$

$f(z) = e^z + z^2 - 2z$ is analytic over the entire complex plane, and i is inside the loop Γ. Thus, we have

$$\int_{\Gamma} \frac{e^z + z^2 - 2z}{(z-i)^3}\, dz = \frac{2\pi i}{2!}\, f''(i) = \pi i(e^i + 2)$$

. . .

The following theorem gives a condition for a function to be analytic in a domain.

Theorem 132. (Morera's Theorem) If f is continuous in a domain D and if

$$\int_{\Gamma} \boldsymbol{f(z)\, dz = 0}$$

for every closed contour Γ in D, then f is analytic in D.

Proof: See the Wikipedia article "Morera's theorem" [98]. ■

4.8 Sequences and Series

4.8.1 Basic Concepts

Consider the following sequence of complex numbers

$$\frac{1}{2i}, \frac{1}{(2i)^2}, \frac{1}{(2i)^3}, \frac{1}{(2i)^4}, \ldots$$

When the denominators are expanded, we get the alternating sequence

$$\frac{1}{2i}, -\frac{1}{2^2}, -\frac{1}{2^3 i}, \frac{1}{2^4}, \frac{1}{2^5 i}, -\frac{1}{2^6}, -\frac{1}{2^7 i}, \ldots$$

Even though the series alternates (2 negative terms, 2 positive terms, …), it does appear to be getting smaller and converging to 0. We can formalize our supposition with the following definition of convergence for a sequence of complex numbers.

A sequence of complex numbers $\{a_n\}_{n=1}^{\infty}$ is said to converge to a if for every $\epsilon > 0$ there exists an N such that $|a - a_n| < \epsilon$ for every $n > N$. In this case, we write $\lim_{n\to\infty} a_n = a$.

Using the above definition, we can prove that the sequence in our example converges to 0. Take $\epsilon > 0$, and consider the inequality

$$\left|0 - \frac{1}{i2^n}\right| = \frac{1}{2^n} < \epsilon \Rightarrow 2^n > \frac{1}{\epsilon}$$

Taking the log base 2 on both sides of the above, we get

$$n \log_2 2 > \log_2\left(\frac{1}{\epsilon}\right) = -\log_2 \epsilon$$

But $\log_2 2 = 1$, and so, the above inequality becomes $n > -\log_2 \epsilon$.

Thus, for a given ϵ, we choose N to be the next integer larger than $-\log_2 \epsilon$. For example, if $\epsilon = .001$, then choose $N = 10$ since $-\log_2(.001) \cong 9.966$.

. . .

Next, we consider infinite series consisting of complex terms, i.e.,

$$\sum_{j=1}^{\infty} a_i\,, \qquad a_j \in \mathbb{C}$$

Let $S_n = \sum_{j=1}^{n} a_j$ be the sum of the first n terms of the series. If $\lim_{n\to\infty} S_n = S$, then the series is said to converge to S. Basically, we have defined convergence of a series in terms of the sequence of partial sums.

Theorem 133. If $\sum_{j=1}^{\infty} a_j$ converges then $\lim_{j\to\infty} a_j = 0$.

Proof: Let $S_n = \sum_{j=1}^{n} a_j$ and note that $a_n = S_n - S_{n-1}$. Since $\sum_{j=1}^{\infty} a_j$ converges, we have that $\lim_{n\to\infty} S_n = S$. This also implies that $\lim_{n\to\infty} S_{n-1} = S$. So, $\lim_{n\to\infty} a_n = \lim_{n\to\infty} S_n - S_{n-1} = \lim_{n\to\infty} S_n - \lim_{n\to\infty} S_{n-1} = S - S = 0$. ■

The following theorem provides a result that is almost identical to a similar result for real numbers.

Theorem 134. The series $\sum_{j=1}^{\infty} a^j$ converges to $\frac{1}{1-a}$ if $|a| < 1$. If $|a| \geq 1$, then the series diverges.

Proof: We have that

$$(1-a)(1 + a + a^2 + \cdots + a^{n-1} + a^n)$$

$$= (1 + a + a^2 + \cdots + a^{n-1} + a^n) - (a + a^2 + \cdots + a^n + a^{n+1})$$

$$= 1 - a^{n+1}$$

which implies

$$\frac{1}{1-a} - (1 + a + a^2 + \cdots + a^{n-1} + a^n) = \frac{a^{n+1}}{1-a}$$

Further, $\lim_{n\to\infty} \frac{a^{n+1}}{1-a} = \frac{a}{1-a} \lim_{n\to\infty} a^n = 0$ since for any $\epsilon > 0$, $|a^n - 0| = |a|^n < \epsilon$ whenever $n > \frac{\ln \epsilon}{\ln |a|}$.

This follows since $|a|^n < \epsilon$ whenever $n \ln |a| < \ln \epsilon$ or whenever $n > \frac{\ln \epsilon}{\ln |a|}$. (Note that $\ln |a|$ is negative since $|a| < 1$ and thus, the direction of the inequality changes when we divide by $\ln |a|$.) Alternatively, we could have appealed to Theorem 133.

Thus,

$$\lim_{n\to\infty}\frac{1}{1-a}-(1+a+a^2+\cdots+a^{n-1}+a^n)=\lim_{n\to\infty}\frac{a^{n+1}}{1-a}=0$$

$$\frac{1}{1-a}=1+a+a^2+\cdots$$

This completes the first part of the proof.

Next, we prove the second part of the theorem. If $|a|\geq 1$, then a^j cannot approach 0 as $j\to\infty$.

Thus, from Theorem 133, $\sum_{j=1}^{\infty}a^j$ cannot converge if $|a|\geq 1$. ■

As an example of the above theorem, we compute the following sum

$$\frac{1}{2i}+\frac{1}{(2i)^2}+\frac{1}{(2i)^3}+\frac{1}{(2i)^4}+\cdots=\frac{1}{2i}\left(1+\frac{1}{2i}+\frac{1}{(2i)^2}+\frac{1}{(2i)^3}+\cdots\right)$$

Taking $a=\frac{1}{2i}$ in the formula of Theorem 134, we have

$$\frac{1}{2i}\sum_{j=1}^{\infty}\frac{1}{(2i)^j}=\frac{1}{2i}\left(\frac{1}{1-\frac{1}{2i}}\right)=\frac{1}{2i}\cdot\frac{2i}{2i-1}=\frac{1}{2i-1}\cdot\frac{-1-2i}{-1-2i}=-\frac{1}{5}(1+2i)$$

. . .

If the terms of a series (from some index and beyond) are bounded by the terms of a series known to be convergent, then the given series also converges. More formally, we have the following theorem which is almost exactly what we have for real series.

Theorem 135. (Comparison Test for Series) If $|a_j|\leq b_j$ for all j greater than some integer N and the real series $\sum_{j=0}^{\infty}b_j$ converges, then so does the complex series $\sum_{j=0}^{\infty}a_j$.

For example, consider the series

$$A=\sum_{j=0}^{\infty}\frac{3i}{(j+2)^j}=3i+i+\frac{3i}{4^2}+\frac{3i}{5^3}+\frac{3i}{6^4}+\cdots$$

Compare this to the convergent series

$$B=3\sum_{j=0}^{\infty}\frac{1}{2^j}=3+\frac{3}{2}+\frac{3}{2^2}+\frac{3}{2^3}+\cdots=3\left(\frac{1}{1-\frac{1}{2}}\right)=6$$

We have that

$$|a_j|=\frac{3}{(j+2)^j}\leq\frac{3}{2^j}=b_j,\qquad\forall j\geq 0$$

By the comparison test, we conclude the series A converges.

. . .

A series $\sum_{j=1}^{\infty} a_j$ is said to be **absolutely convergent** if $\sum_{j=1}^{\infty}|a_j|$ converges. It follows from the comparison test that an absolutely convergent series is convergent. The converse is not true. For example, the series $\sum_{j=1}^{\infty} i\frac{(-1)^n}{n} = i\sum_{j=1}^{\infty}\frac{(-1)^n}{n} = i\ln 2$ converges but the series $\sum_{j=1}^{\infty}\left|i\frac{(-1)^n}{n}\right| = \sum_{j=1}^{\infty}\frac{1}{n}$ diverges. (Note that $\sum_{j=1}^{\infty}\frac{(-1)^n}{n}$ is known as the alternating harmonic series.)

There is also a version of the ratio test for series converges for complex series.

Theorem 136. (Ratio Test) If for a series $\sum_{j=1}^{\infty} a_j$ we have that $\lim_{j\to\infty}\left|\frac{a_{j+1}}{a_j}\right| = A$ then the series converges if $A < 1$, diverges if $A > 1$ (with no conclusion if $A = 1$).

Proof: Assume $A < 1$ and select real number r such that $A < r < 1$. This implies that $|a_{n+1}| < r|a_n|$ for all n greater than some sufficiently large number N. Thus, $|a_{n+j}| < r^j|a_n|$ for all $n > N$ and $j > 0$ and so,

$$\sum_{j=N+1}^{\infty}|a_j| = \sum_{j=1}^{\infty}|a_{N+j}| < \sum_{j=1}^{\infty} r^j|a_N| = |a_N|\sum_{j=1}^{\infty} r^j = |a_N|\frac{r}{1-r} < \infty$$

So, the series converges absolutely but as we noted above, this implies the series converges.

If $A > 1$, then the ratio $\left|\frac{a_{n+1}}{a_n}\right|$ will eventually be greater than 1, i.e., there exists an integer N such that $\left|\frac{a_{n+1}}{a_n}\right| > 1$ for all $n \geq N$. For this value of N, we have that $|a_{N+k}| > |a_N|$ for all $k \geq 1$. So, $\lim_{n\to\infty} a_n \neq 0$ and by Theorem 133, $\sum_{j=1}^{\infty} a_j$ diverges. ■

For example, take the series $\sum_{j=0}^{\infty}\frac{(3i)^j}{j!}$. Applying the ratio test, we have

$$A = \left|\frac{a_{j+1}}{a_j}\right| = \frac{|(3i)^{j+1}|}{(j+1)!}\frac{j!}{|(3i)^j|} = \frac{3^{j+1}\,|i|^{j+1}}{j\,3^j\,|i|^j} = \frac{3}{j}$$

Since $\lim_{j\to\infty} A = \lim_{j\to\infty}\frac{3}{j} = 0 < 1$, the series converges by the ratio test.

. . .

Convergence can be defined for a sequence of complex functions as follows.

A sequence of complex functions $\{f_n(z)\}_{n=1}^{\infty}$ is said to **converge pointwise** to a if for every $\epsilon > 0$ there exists an $N(\epsilon, z)$ such that $|a - a_n| < \epsilon$ for every $n > N(\epsilon, z)$. In this case, we write $\lim_{n\to\infty} f_n(z) = a$.

It is important to note the N depends on both ϵ and z. If N only depends on ϵ and the limit exists, then the function is said to **converge uniformly**. In terms of series, $\sum_{j=0}^{\infty} f_j(z)$ converges uniformly to some function $f(z)$ if the sequence of partial sums $\{S_n = \sum_{j=0}^{n} f_j(z)\}$ converges uniformly to $f(z)$.

Consider the following series, where $c \in \mathbb{C}$ is a constant,

$$\sum_{j=0}^{\infty} (cz)^j$$

In the proof of Theorem 134, we derived the following formula

$$\frac{1}{1-a} - (1 + a + a^2 + \cdots + a^{n-1} + a^n) = \frac{a^{n+1}}{1-a}$$

Letting $a = cz$, we have

$$\frac{1}{1-cz} - (1 + cz + (cz)^2 + \cdots + (cz)^{n-1} + (cz)^n) = \frac{(cz)^{n+1}}{1-cz} \quad \text{(Equation 1)}$$

So, when $|cz| < 1$ or equivalently $|z| < \frac{1}{|c|}$, the series converges to $\frac{1}{1-cz}$.

To prove the uniform convergence of the above series, consider the closed disk $|z| \le r$ for $r < \frac{1}{|c|}$. Under this assumption, we have the following

$$|(cz)^{n+1}| \le (r|c|)^{n+1}$$

$$|cz| \le r|c| < 1$$

$$|1 - cz| \ge \big|1 - |cz|\big| \ge \big|1 - r|c|\big|$$

From Equation 1 above, we have

$$\left|\frac{(cz)^{n+1}}{1-cz}\right| \le \frac{(r|c|)^{n+1}}{\big|1 - r|c|\big|}, \qquad |z| \le r$$

The above inequality, which is a bound on the remainder of $\frac{1}{1-cz}$ minus the partial sum $\sum_{j=0}^{n}(cz)^j$, can be made less than any $\epsilon > 0$ for n sufficiently large, **independent of z**. Thus, the series $\sum_{j=0}^{\infty}(cz)^j$ converges uniformly on the closed disk $|z| \le r < \frac{1}{|c|}$.

Some exercises to try:

- Show that the series $\sum_{n=1}^{\infty}\left[\frac{2}{n^3}-i^n\right]$ diverges. **Hint**: consider the real and imaginary parts separately.
- Determine whether $\sum_{n=1}^{\infty}\left(\frac{1}{2}+i\right)^n$ converges. **Hint**: $\left|\frac{1}{2}+i\right|=\sqrt{\frac{1}{4}+1}=\sqrt{\frac{5}{4}}>1$.
- Determine whether $\sum_{n=1}^{\infty}\frac{n(\,i^n)}{3n+1}$ converges. **Hint**: As $n\to\infty$, $\frac{n}{3n+1}\to\frac{1}{3}$ and i^n alternates around 4 values, i.e., $1, i, -1, -i$.

4.8.2 Taylor, Maclaurin and Laurent Series

The **Taylor series** of a complex or real function is an infinite sum that is expressed in terms of the function's derivatives at a single point. We have the following theorem concerning the Taylor series for an analytic function.

Theorem 137. If function $f(z)$ is analytic in the disk $|z-z_0|<R$, then

$$f(z)=\sum_{n=0}^{\infty}a_n(z-z_0)^n,\qquad |z-z_o|<R$$

$$a_n=\frac{f^{(n)}(z_0)}{n!}$$

The series converges to $f(z)$ for z in the stated disk, and the convergence is uniform in any closed disk $|z-z_0|\le r<R$. The series representation is unique within the given disk.

(Note that $f^{(n)}$ denotes the n^{th} derivative of f.)

Proof: See Sections 58 and 66 of "Complex Variables and Applications" [91]. ■

When $z_0=0$ in the above formula, the series is known as a **Maclaurin series**.

As an example, consider $f(z)=e^z$ which is analytic over the entire complex plane. We have that

$$a_n=\frac{f^{(n)}(0)}{n!}=\frac{e^0}{n!}=\frac{1}{n!}$$

Thus, the Maclaurin series for e^z is

$$\sum_{n=0}^{\infty}\frac{z^n}{n!}$$

We can multiply a Taylor series by a constant to get another valid Taylor series, and we can also add and subtract Taylor series.

Theorem 138. Given Talyor series $\boldsymbol{f(z) = \sum_{j=0}^{\infty} a_j (z - z_0)^j}$ ***and*** $\boldsymbol{g(z) = \sum_{j=0}^{\infty} b_j (z - z_0)^j}$ ***where both*** $\boldsymbol{f}$ ***and*** $\boldsymbol{g}$ ***are analytic within an open disk about*** $\boldsymbol{z_0}$***. Then the Taylor series for*** $\boldsymbol{cf(z)}$ ***with*** $\boldsymbol{c \in \mathbb{C}}$ ***is*** $\boldsymbol{\sum_{j=0}^{\infty} c a_j (z - z_0)^j}$ ***and the Taylor series for*** $\boldsymbol{f(z) \pm g(z)}$ ***is*** $\boldsymbol{\sum_{j=0}^{\infty} (a_j \pm b_j)(z - z_0)^j}$***.***

Proof: This follows from the fact that a linear combination of two analytic functions is analytic (assuming the same domain of analyticity). ■

For example, consider $\sin z$ which, as we have seen, is defined as the difference of two analytic functions, i.e., $\sin z = \frac{e^{iz}}{2i} - \frac{e^{(-iz)}}{2i}$. Applying Theorem 138 and our previous result for the Maclaurin series for e^z, we have

$$sin\, z = \frac{1}{2i} \sum_{n=0}^{\infty} \frac{(iz)^n}{n!} - \sum_{n=0}^{\infty} \frac{(-iz)^n}{n!} = \frac{1}{2i} \sum_{n=0}^{\infty} [1 - (-1)^n] \frac{(i^n z^n)}{n!}, \qquad |z| < \infty$$

Since $1 - (-1)^n = 0$ when n is even, we only need to consider the odd terms in the above (which is accomplished by replacing n by $2n + 1$).

$$sin\, z = \frac{1}{2i} \sum_{n=0}^{\infty} [1 - (-1)^{2n+1}] \frac{(i^{2n+1} z^{2n+1})}{(2n+1)!} = \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{(2n+1)!}, \qquad |z| < \infty$$

When going to the final result in the above, we used the facts that

$$1 - (-1)^{2n+1} = 2, \qquad i^{2n+1} = (i^2)^n i = (-1)^n i$$

. . .

Consider the principal branch of $f(z) = \log z$ in the domain $|z - 1| < 1$. In terms of derivatives, we have

$$\frac{d}{dz} \log z = \frac{1}{z}, \frac{d^2}{dz^2} \log z = -\frac{1}{z^2}, \frac{d^3}{dz^3} \log z = -\frac{2!}{z^3}, \ldots, \frac{d^n}{dz^n} \log z = (-1)^{n+1} (n-1)! \frac{1}{z^n}, \ldots$$

Applying Theorem 137, we have

$$a_n = \frac{f^{(n)}(1)}{n!} = \frac{(-1)^{n+1}}{n}$$

So, the Taylor series for $\log z$ about the point 1 is

$$\log z = \sum_{n=0}^{\infty} \frac{(-1)^{n+1} (z-1)^n}{n}$$

. . .

Theorem 139. If a function f is analytic at point z_0, one can determine the Taylor series for $f'(z)$ by taking the derivative of the Taylor series of $f(z)$ on a term-by-term basis. The resulting series converges on the same open disk as the Taylor series of $f(z)$.

Proof: Noting that the n^{th} derivative of f' is the $(n+1)^{st}$ derivative of f, we have that the Taylor series for f' is

$$f'(z_0) + f''(z_0)(z - z_0) + \frac{f'''(z_0)}{2!}(z - z_0) + \cdots$$

If we take the derivative of the Taylor series for f, we get exactly the same expression as above.

Application of Theorem 137 to $f'(z)$ implies that the above expression converges in the largest open disk around z_0 over which $f'(z)$ is analytic and based on Theorem 131, f' is analytic wherever f is analytic. ■

We can apply Theorem 139 to the Taylor series that we derived for $\sin z$ to get the Taylor series for $\cos z$.

$$\cos z = \frac{d}{dz}\sin z = \frac{d}{dz}\sum_{n=0}^{\infty}\frac{(-1)^n z^{2n+1}}{(2n+1)!}$$

$$= \sum_{n=0}^{\infty}\frac{d}{dz}\left[\frac{(-1)^n z^{2n+1}}{(2n+1)!}\right] = \sum_{n=0}^{\infty}\frac{(-1)^n z^{2n}}{(2n)!}, \qquad |z| < \infty$$

. . .

Given the identity $\sinh z = -i\sin(iz)$, we can compute the Taylor series of $\sinh z$ by appropriate substitution into the Taylor of $\sin z$.

$$\sinh z = -i\sin(iz) = -i\sum_{n=0}^{\infty}\frac{(-1)^n (iz)^{2n+1}}{(2n+1)!}$$

$$= \sum_{n=0}^{\infty}\frac{(-1)^{n+1}\, i^{2n+2}\, z^{2n+1}}{(2n+1)!} = \sum_{n=0}^{\infty}\frac{z^{2n+1}}{(2n+1)!}, \qquad |z| < \infty$$

For the last simplification in the above, we used the fact that

$$(-1)^{n+1}\, i^{2n+2} = (-1)^{n+1}\,(i^2)^{n+1} = (-1)^{n+1}\,(-1)^{n+1} = (-1)^{2n+2} = 1$$

Using a similar idea, we can derive the Taylor series for $\cosh z$ from the identity $\cosh z = \cos(iz)$ and the Taylor series for $\cos z$.

$$\cosh z = \sum_{n=0}^{\infty}\frac{z^{2n}}{(2n)!}, \qquad |z| < \infty$$

. . .

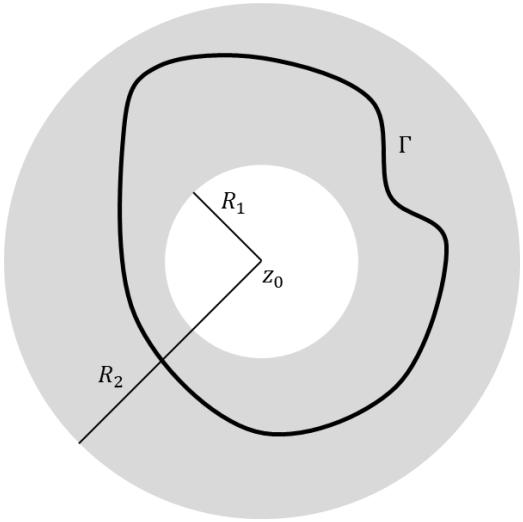
In cases where a function is not defined at one or more points within a given disk, the Taylor series does not apply. However, it may still be possible in such cases to define what is known as a **Laurent series**. The Laurent series for a function is defined over an annulus where the function is analytic.

Theorem 140. If a function f is analytic in an annular domain $R_1 < |z - z_0| < R_2$ and Γ is a positively oriented simple closed curved within the annulus, then f has a unique series representation of the following form:

$$f(z) = \sum_{n=-\infty}^{\infty} c_n (z - z_0)^n, \qquad R_1 < |z - z_0| < R_1$$

$$c_n = \frac{1}{2\pi i} \int_{\Gamma} \frac{f(z)}{(z - z_0)^{n+1}}\, dz, \qquad n = 0, \pm 1, \pm 2, \ldots$$

The annulus (gray region), point z_0 and contour Γ associated with the theorem are shown in the figure below.



Proof: See Sections 61 and 66 of "Complex Variables and Applications" [91].

The coefficients in a Laurent series are typically found by means other than through the use of their integral representations (as we shall see in the following examples.

The Laurent series from $e^{\frac{1}{z}}$ in the annulus $0 < |z| < \infty$ can be determined by substituting $\frac{1}{z}$ into the Taylor series for e^z.

$$e^{\frac{1}{z}} = 1 + \frac{1}{z} + \frac{1}{2!}\frac{1}{z^2} + \frac{1}{3!}\frac{1}{z^3} + \cdots$$

$$\ldots$$

The function $f(z) = \frac{z^3+3z^2-2z+7}{z}$ can be expressed as the Laurent series

$$z^2 + 3z - 2 + \frac{7}{z}, \qquad 0 < |z| < \infty$$

In this case, the region of analyticity is the entire complex plane minus the point $z = 0$. Such a region, i.e., a disk with its center removed, is called a punctured disk. In this particular case, the radius of the disk is infinite.

. . .

Consider the function $f(z) = \frac{2z}{z^2+1}$. In terms of partial fractions, this can be written as

$$f(z) = \frac{1}{z-i} + \frac{1}{z+i}$$

Next, we determine the Taylor series representation of $g(z) = \frac{1}{z+i}$ at $z_0 = i$, noting that $\frac{1}{z+i}$ is analytic in the open disk $|z - i| < 2$.

$$\frac{1}{z+i} = \sum_{n=0}^{\infty} a_n (z-i)^n = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2i)^{n+1}} (z-i)^n = \frac{1}{2i} \sum_{n=0}^{\infty} (-1)^n \left(\frac{z-i}{2i}\right)^n, \qquad |z-i| < 2$$

where

$$g^{(n)}(z) = \frac{(-1)^n\, n!}{(z+i)^{n+1}}, \qquad a_n = \frac{g^{(n)}(i)}{n!} = \frac{(-1)^n}{(2i)^{n+1}}$$

Thus, the Laurent series for $f(z)$ about $z_0 = i$ is

$$\frac{1}{z-i} + \frac{1}{2i} \sum_{n=0}^{\infty} (-1)^n \left(\frac{z-i}{2i}\right)^n = \frac{1}{z-i} + \sum_{n=0}^{\infty} \frac{-i^{n+1}\,(z-i)^n}{2^{n+1}}, \qquad 0 < |z-i| < 2$$

In the above, we made use of the following simplification:

$$\frac{1}{i} \cdot \frac{(-1)^n}{i^n} = \frac{1}{i} \cdot \left(-\frac{1}{i}\right)^n = \frac{1}{i} \cdot i^n = (-i) i^n = -i^{n+1}$$

Alternatively, we could have used a similar approach to develop the Laurent series for $f(z)$ about $z_0 = -i$. This is left as an exercise for the reader. You can check your answer using Wolfram Alpha by entering the command

laurent series for 2z/(z^2+1) about z=-i

. . .

The function $f(z) = \frac{z+1}{z^3(z^2+1)}$ has singularities at $z = -i, 0, i$. To compute the Laurent series about $z_0 = 0$ in the region $0 < |z| < 1$, we first expand the geometric series for $\frac{1}{1+z^2}$

$$\frac{1}{1+z^2} = \frac{1}{1-(z^2)} = 1 - z^2 + z^4 - z^6 \pm \cdots, \qquad |z| < 1$$

Multiplying the above by $z + 1$ we get

$$1 + z - z^2 - z^3 + z^4 + z^5 - z^6 - z^7 \pm \cdots$$

Multiplying by $\frac{1}{z^3}$, we get the Laurent series for $f(z)$.

$$f(z) = \frac{1}{z^3} + \frac{1}{z^2} - \frac{1}{z} - 1 + z + z^2 - z^3 - z^4 \pm \cdots, 0 < |z| < 1$$

. . .

The Laurent series depends on the selected region. For example, consider the function

$$f(z) = \frac{-1}{z^3(z-1)}$$

In the region $0 < |z| < 1$, we have the Laurent series

$$\frac{-1}{z^3(z-1)} = \frac{1}{z^3}\left(\frac{1}{1-z}\right) = \frac{1}{z^3}(1 + z + z^2 + z^3 + \cdots) = \frac{1}{z^3} + \frac{1}{z^2} + \frac{1}{z} + 1 + z + z^2 + \cdots$$

In the region $1 < |z| < \infty$, we need to modify our approach so that the associated geometric series converges. Note that $1 < |z|$ implies $\left|\frac{1}{z}\right| < 1$ and thus, the geometric series $\frac{1}{1-\frac{1}{z}}$ converges. So, in the region $1 < |z| < \infty$, we have the Laurent series

$$\frac{-1}{z^3(z-1)} = -\frac{1}{z^3} \cdot \frac{1}{z}\left(\frac{1}{1-\frac{1}{z}}\right) = -\frac{1}{z^4}\left(1 + \frac{1}{z} + \frac{1}{z^2} + \cdots\right) = -\frac{1}{z^4} - \frac{1}{z^5} - \frac{1}{z^6} - \cdots$$

. . .

Applying partial fraction decomposition to the function $f(z) = \frac{z}{z^2-3z+2}$, we get

$$f(z) = \frac{z}{(z-1)(z-2)} = \frac{2}{z-2} - \frac{1}{z-1}$$

The two singularities, i.e., $z = 1,2$, leave us with three domains to consider when construction a series representation for $f(z)$.

In the domain $|z| < 1$, we have that the Maclaurin series for $\frac{1}{1-\frac{z}{2}}$ converges for $|z/2| < 1$ or $|z| < 2$, and the Maclaurin series $\frac{1}{1-z}$ converges for $|z| < 1$. Thus, the following converges for the intersection of $|z| < 2$ and $|z| < 1$, which is $|z| < 1$.

$$f(z) = \frac{2}{z-2} - \frac{1}{z-1} = -\frac{1}{1-\frac{z}{2}} + \frac{1}{1-z}$$

$$= -\sum_{n=0}^{\infty} \frac{z^n}{2^n} + \sum_{n=0}^{\infty} z^n = \sum_{n=0}^{\infty} \left(1 - \frac{1}{2^n}\right) z^n$$

For $1 < |z| < 2$, we have the following Laurent series for $f(z)$ which converges for the intersection of $\left|\frac{1}{z}\right| < 1$ and $\left|\frac{z}{2}\right| < 1$, i.e., $1 < |z| < 2$.

$$f(z) = -\frac{1}{z} \cdot \frac{1}{1-\frac{1}{z}} - \frac{1}{1-\frac{z}{2}}$$

$$= -\frac{1}{z} \sum_{n=0}^{\infty} \frac{1}{z^n} - \sum_{n=0}^{\infty} \frac{z^n}{2^n} = -\left(\sum_{n=1}^{\infty} \frac{1}{z^n} + \sum_{n=0}^{\infty} \frac{z^n}{2^n}\right)$$

For $|z| > 2$, we have the following Laurent series for $f(z)$ which converges for the intersection of $\left|\frac{1}{z}\right| < 1$ and $\left|\frac{2}{z}\right| < 1$, i.e., $|z| > 2$.

$$f(z) = -\frac{1}{z} \cdot \frac{1}{1-\frac{1}{z}} + \frac{2}{z} \cdot \frac{1}{1-\frac{2}{z}}$$

$$= -\frac{1}{z} \sum_{n=0}^{\infty} \frac{1}{z^n} + \frac{2}{z} \sum_{n=0}^{\infty} \frac{2^n}{z^n} = -\sum_{n=0}^{\infty} \frac{1}{z^{n+1}} + \sum_{n=0}^{\infty} \frac{2^{n+1}}{z^{n+1}} = \sum_{n=1}^{\infty} \frac{2^n - 1}{z^n}$$

4.9 Classifications Zeros and Singularities

The Taylor and Laurent series can be used to study the behavior of an analytic function near its zeros and isolated singularities. As we have seen, a zero (or root) of a function f is a point z such that $f(z) = 0$. An isolated singularity of function f is a point z_0 such that f is analytic in some punctured disk $0 < |z - z_0| < r$ but not analytic at z_0 itself. For example, the function

$$f(z) = \frac{\sin z}{z}$$

has an isolated singularity at $z = 0$, and has zeros at $z = \pm n\pi, n = 1,2,3, \ldots$

Consider a function f that is analytic in a domain D containing a point z_0. Further, assume $f(z_0) = f'(z_0) = f''(z_0) = \cdots = f^{(n-1)}(z_0) = 0$ but $f^{(n)}(z_0) = 0$, then f is said to have a **zero of order n** at point z_0. In this case, the Taylor series around z_0 has the form

$$f(z) = a_n(z - z_0)^n + a_{n+1}(z - z_0)^{n+1} + a_{n+2}(z - z_0)^{n+2} + \cdots$$

$$= (z - z_0)^n g(z)$$

where

$$g(z) = a_n + a_{n+1}(z - z_0) + a_{n+2}(z - z_0)^2 + \cdots$$

The function $g(z)$ converges wherever the series for $f(z)$ converges, since for any given point, the functions are just a constant multiple of each other. So, $g(z)$ is analytic in the domain D about z_0 and $g(z_0) \neq 0$.

Conversely, any function f of the form $f(z) = (z - z)^n g(z)$, as described above, must have a zero of order n. Thus, we have proven the following theorem which is basically an extension of a previous result for polynomials.

Theorem 141. Given a function f that is analytic in a domain about point z_0. f has a zero of order n at z_0 if and only if f can be written in the form $f(z) = (z - z_0)^n g(z)$ where $g(z)$ is analytic in said domain and $g(z_0) \neq 0$.

Zeros of non-constant analytic functions are isolated, as demonstrated in the proof of the following theorem.

Theorem 142. If $f(z_0) = 0$ and f is analytic in a disk about z_0, then f is identically zero in the disk about z_0 or there is a punctured disk about z_0 (i.e., disk with point z_0 removed) in which f has no zeros.

Proof: By Theorem 137, the Taylor series $f(z) = \sum_{n=0}^{\infty} a_n(z - z_0)^n$ converges in some open disk $|z - z_0| < R$.

- If $a_n = 0$ for all n, then $f(z)$ must be identically zero in the disk.
- If all the a_n are not zero, let $k \geq 1$ be the smallest subscript such that $a_k \neq 0$, i.e., $f(z)$ has a zero of order k at z_0. By Theorem 141, $f(z)$ can be represented in the form $(z - z_0)^k g(z)$ where $g(z_0) \neq 0$. Since $g(z)$ is analytic (and thus continuous) at z_0, there exists a disk $|z - z_0| < r \leq R$ over which g is nonzero, and so, f is nonzero in the punctured disk $0 < |z - z_0| < r$. ■

If f is nonconstant, analytic and equal to zero at z_0, the order of the zero must be a positive integer. However, if f is not analytic, the order of its zero is not necessarily an integer (e.g., $z^{\frac{1}{2}}$ has zero of order $\frac{1}{2}$) or even defined (e.g., $|z|$).

A point z_0 is an isolated singularity of a function f if f is analytic in a punctured disk $0 < |z - z_0| < R$ but not analytic at z_0. For example, $\tan\frac{\pi z}{2}$ has an isolated singularity at each odd integer value of z.

The isolated singularities of a function f can be classified as follows. The various c_n terms refer to the coefficients of the associated Laurent series.

- If $c_n = 0$ for all $n < 0$, then z_0 is said to be a **removable singularity** of f. In such cases, the $\lim_{z \to z_0} f(z) = L$ exists. Thus, we can set $f(z_0) = L$, and effectively remove the singularity.
- If $c_{-k} \neq 0$ for some positive integer k but $c_n = 0$ for all $n < -k$, then z_0 is said to be a **pole of order k** of f. A pole of order 1 is known as a **simple pole**.
- If $c_n \neq 0$ for an infinite number of negative values for n, then z_0 is said to be an **essential singularity** of f. This is equivalent to the condition that neither $\lim_{z \to z_0} f(z)$ nor $\lim_{z \to z_0} \frac{1}{f(z)}$ exists.

In the case of a removable singularity, the associated Taylor series takes the form

$$f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n = a_0 + a_1 (z - z_0) + a_2 (z - z_0)^2 + \cdots$$

We define $f(z_0)$ to be a_0, and thus, remove the singularity.

. . .

For example, $f(z) = \frac{\sin z}{z}$ has a singularity at $z = 0$. However, it has Taylor series

$$\frac{1}{z}\left(z - \frac{z^3}{3!} + \frac{z^5}{5!} \pm \cdots\right) = 1 - \frac{z^2}{3!} + \frac{z^4}{5!} \pm \cdots$$

So, we can remove the singularity by setting $f(0)$ to 1.

The function $f(z) = \frac{e^z - 1}{z}$ has a singularity at $z = 0$. However, it has Taylor series

$$\frac{1}{z}\left[\left(1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots\right) - 1\right] = 1 + \frac{z}{2!} + \frac{z^2}{3!} + \frac{z^3}{4!} + \cdots$$

So, we can remove the singularity by setting $f(0)$ to 1.

The function $f(z) = \frac{z^2 - 4}{z - 2}$ has a singularity at $z = 2$. By factoring the numerator, we can easily get its Taylor series, i.e., $z + 2$. The singularity can be removed by setting $f(2)$ to 4.

. . .

The function $f(z) = \frac{e^z - 1}{z^4}$ has a singularity at $z = 0$. Its Laurent series is

$$\frac{e^z - 1}{z^4} = \frac{1}{z^4}\left[\left(1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots\right) - 1\right] = \frac{1}{z^3} + \frac{1}{2!}\frac{1}{z^2} + \frac{1}{3!}\frac{1}{z} + \frac{1}{4!} + \frac{1}{5!}z + \cdots$$

Thus, the singularity is a pole of order 3.

. . .

The function $f(z) = e^{\frac{1}{z}}$ has an essential singularity at $z = 0$ as one can see by examining the associated Laurent series

$$e^{\frac{1}{z}} = 1 + \frac{1}{z} + \frac{1}{2!}\frac{1}{z^2} + \frac{1}{3!}\frac{1}{z^3} + \cdots$$

The functions $\cos\frac{1}{z}$ and $\sin\frac{1}{z}$ also have essentially singularities at $z = 0$.

. . .

The following theorem summarizes some concepts concerning removable singularities.

Theorem 143. If a function f has a removable singularity at z_0, then

- ***i. $f(z)$ is bounded in some punctured disk about z_0.***
- ***ii. $f(z)$ has a finite limit as $z \to z_0$.***
- ***iii. $f(z_0)$ can be redefined to remove the singularity and make f analytic at z_0.***

Conversely, if a function f is bounded in some punctured disk about an isolated singularity z_0, then the singularity is removable.

Picard's little theorem and Picard's great theorem are related theorems about the codomain of an analytic function.

Theorem 144. (Picard's Little Theorem) If a function f is entire and non-constant, then the set of values assumed by f is either the whole complex plane or the plane minus a single point.

The "single point" exception is needed in the above theorem. For example, e^z is entire and non-constant, but never takes the value 0.

Theorem 145. (Picard's Great Theorem) If an analytic function f has an essential singularity at a point z_0, then on any punctured disk about z_0 , f takes on all possible complex values, with at most a single exception, infinitely often.

Proof: For proofs of both theorems, see the Wikipedia article "Picard theorem" [99]. ■

The following theorem summarizes the equivalent conditions for the various types of singularities.

Theorem 146. For a singularity z_0 of function f, the following equivalent conditions hold true for the various types of singularities.

- ***i. z_0 is a removable singularity $\Leftrightarrow |f(z)|$ is bounded in some punctured neighborhood about $z_0 \Leftrightarrow \lim_{z \to z_0} f(z)$ exists $\Leftrightarrow f$ can be redefined to be analytic at z_0.***
- ***ii. z_0 is a pole of order $m \Leftrightarrow \lim_{z \to z_0} |f(z)| = \infty \Leftrightarrow f(z)$ can be written in the form $\frac{g(z)}{(z-z_0)^m}$ in some punctured neighborhood about z_0, where $g(z)$ is an analytic function such that $g(z_0) \neq 0$.***

iii. ***z_0 is an essential singularity $\Leftrightarrow |f(z)|$ is neither bounded in any neighborhood of z_0 nor goes to infinity as $z \to z_0 \Leftrightarrow f(z)$ assumes every complex number in every neighborhood of z_0, with one possible exception.***

Proof: See Section 5.6 of "Fundamentals of Complex Analysis [95]. ■

4.10 Residue Theory

4.10.1 Concepts

Recall the example associated with Figure 73. The example was that of computing a contour integral about two simple singularities. In what follows, we will develop a method for computing the integral of functions that are analytic within a given contour except for a finite number of singularities within the contour.

In general, consider a function $f(z)$ which is analytic on and within a simple closed positively oriented contour Γ except for a finite number of isolated singularities within Γ, see Figure 74.

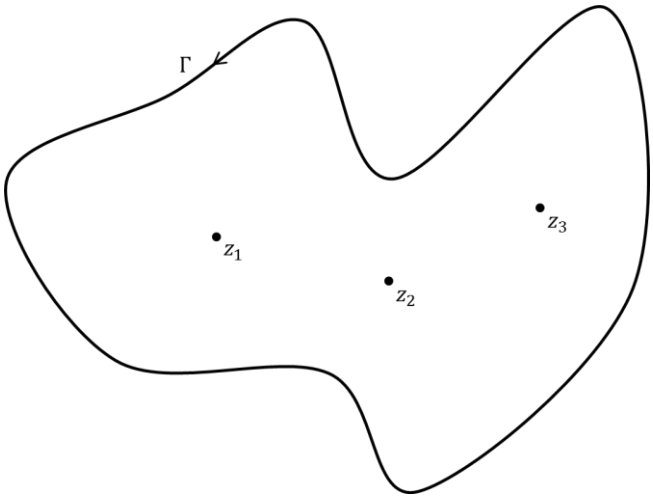


Figure 74. Function with several isolated singularities within a contour

As we did with the example related to Figure 73, we can reduce the problem to computing the integral around several circles (each about one of the singularities). First, we divide Γ into two contours, i.e., Γ_1 which goes from w_2 to w_1, and Γ_2 which goes from w_1 to w_2 (see Figure 75). We have that

$$\int_{\Gamma} f(z)\, dz = \int_{\Gamma_1} f(z)\, dz + \int_{\Gamma_2} f(z)\, dz$$

Next, we create contour Ω from w_2 to w_1. By Theorem 127 (deformation invariance theorem),

$$\int_{\Gamma_1} f(z)\, dz = \int_{\Omega} f(z)\, dz$$

Similarly, we have that

$$\int_{\Gamma_2} f(z)\,dz = \int_{\Phi} f(z)\,dz$$

Regarding contours Ω and Φ, the integral over contour ω_7 cancels ϕ_1, ω_5 cancels ϕ_3, ω_3 cancels ϕ_5, and ω_1 cancels ϕ_7. The remaining contours can be combined to form three circular contours, i.e., C_1, C_2 and C_3 as labeled in Figure 75.

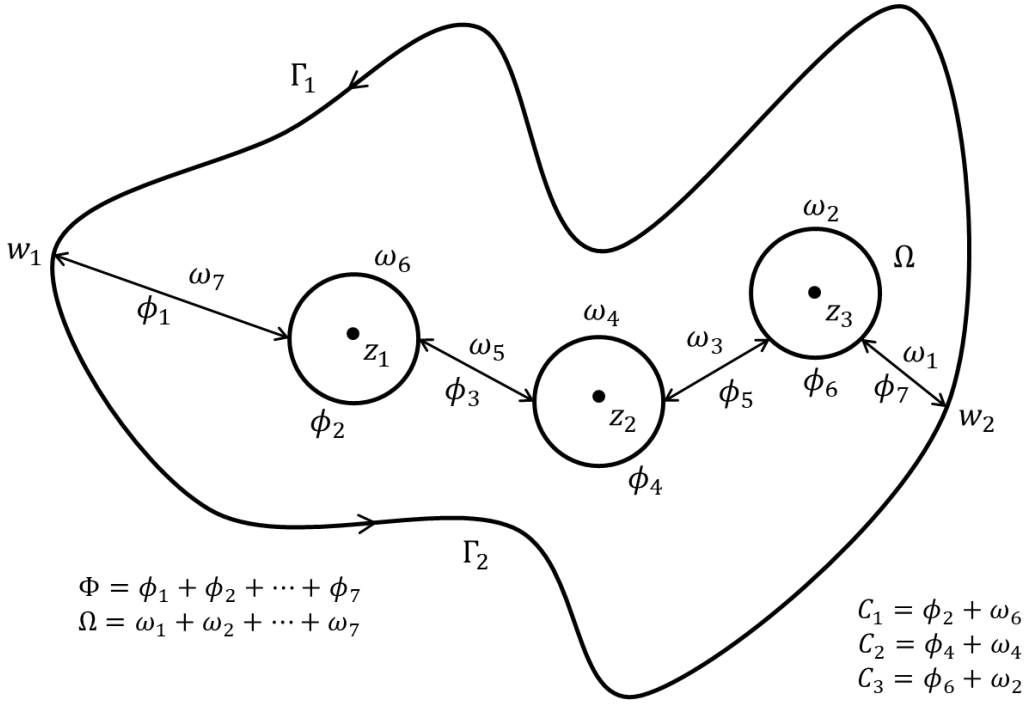


Figure 75. Reduce problem to several circular contours

So, we have reduced the problem as follows:

$$\int_{\Gamma} f(z)\,dz = \int_{\Gamma_1} f(z)\,dz + \int_{\Gamma_2} f(z)\,dz = \int_{\Omega} f(z)\,dz + \int_{\Phi} f(z)\,dz$$

$$= \int_{C_1} f(z)\,dz + \int_{C_2} f(z)\,dz + \int_{C_3} f(z)\,dz$$

There is nothing special about 3 singularities. We can generalize the above method for n isolated singularities of the function $f(z)$, i.e.,

$$\int_{\Gamma} f(z)\,dz = \sum_{j=1}^{n} \int_{C_j} f(z)\,dz$$

where $z_1, z_2, \ldots, z_n$ are isolated singularities of $f(z)$ and C_j is a positively oriented circular contour about singularity about z_j. Further, we assume circular contours are disjoint and are contained within Γ. The above result is known as **Cauchy's residue theorem**.

At this point, we are faced with the task of computing a series of integrals (each about an isolated singularity). We will need a different technique depending on whether we have a removable singularity, a pole or an essential singularity.

Consider a function $f(z)$ which is analytic in a punctured disk about a singularity z_0. Let C be the circle comprising the boundary of the punctured disk. By Theorem 140, $f(z)$ has Laurent series given by

$$f(z) = \sum_{n=-\infty}^{\infty} c_n (z - z_0)^n$$

So,

$$\int_C f(z)\, dz = \sum_{n=-\infty}^{\infty} c_n \int_C (z - z_0)^n \, dz$$

We've already computed the above integrals in an example from Section 4.7.2. Recall that the result was

$$\int_C (z - z_0)^n \, dz = \begin{cases} 0, & n \neq -1 \\ 2\pi i, & n = -1 \end{cases}$$

Thus,

$$\int_C f(z)\, dz = 2\pi i \, c_{-1}$$

In general, the term c_{-1} is known as the **residue** of f at z_0 and is denoted by $Res(f; z_0)$.

In the case of an isolated removable singularity z_0, all the coefficients of negative powers of $(z - z_0)$ in the Laurent series are zero. So, $Res(f; z_0) = 0$ in this case. For example, take $f(z) = \frac{\sin z}{z}$ and $z_0 = 0$. Without having to compute the integral about a disk centered at $z_0 = 0$, we know that $Res\left(\frac{\sin z}{z}; 0\right) = 0$.

In the case function f has a simple pole at z_0, the Laurent expansion about z_0 has the form

$$f(z) = \frac{c_{-1}}{z - z_0} + c_0 + c_1(z - z_0) + c_2(z - z_0)^2 + \cdots$$

which implies

$$(z - z_0)f(z) = c_{-1} + (z - z_0)[c_1(z - z_0) + c_2(z - z_0)^2 + \cdots]$$

Thus, $\lim_{z \to z_0} (z - z_0) f(z) = c_{-1} = Res(f; z_0)$.

For example, the function $f(z) = \frac{\sin z}{(z-1)(z+2)}$ has simple poles at $z = 1$ and $z = -2$ (by Part ii of Theorem 146). Using the above result for the residue of a simple pole, we have

$$Res(f; 1) = \lim_{z \to 1} (z - 1) f(z) = \lim_{z \to 1} \frac{\sin z}{z + 2} = \frac{\sin 1}{3}$$

$$Res(f; -2) = \lim_{z \to -2} (z + 2) f(z) = \lim_{z \to -2} \frac{\sin z}{z - 1} = -\frac{\sin(-2)}{3}$$

. . .

Consider the function $f(z) = \frac{p(z)}{q(z)}$ where $p(z)$ and $q(z)$ are analytic at z_0, $q(z)$ has a simple zero at z_0, and $p(z_0) \neq 0$. So, $q(z) = (z - z_0)h(z)$ where $h(z_0) \neq 0$. We can write $f(z)$ as

$$f(z) = \frac{p(z)/h(z)}{(z - z_0)}$$

By Part ii of Theorem 146, $f(z)$ has a simple pole at z_0 and thus, we can apply our residue formula for simple poles, i.e.,

$$Res(f; z_0) = \lim_{z \to z_0} (z - z_0) \frac{p(z)}{q(z)} = \lim_{z \to z_0} \frac{p(z)}{\frac{q(z) - q(z_0)}{z - z_0}} = \frac{p(z_0)}{q'(z_0)}$$

For example, consider $f(z) = \tan z = \frac{\sin z}{\cos z}$. We have that $\sin z$ and $\cos z$ are entire (i.e., analytic over the entire complex plane), $\cos z = 0$ at $z = \frac{\pi}{2} \pm n\pi, n = 0,1,2, \ldots$ and $\sin z \neq 0$ for any point z where $\cos z = 0$. So, $\tan z$ has simple poles at $z = \frac{\pi}{2} \pm n\pi, n = 0,1,2, \ldots$ and we meet the conditions of the formula for the residue of functions of the form $\frac{p(z)}{q(z)}$. Noting the $\frac{d}{dz} \cos z = -\sin z$, and using the formula that we derived above, we have that

$$Res\left(\tan z, \frac{\pi}{2} \pm n\pi\right) = \frac{\sin\left(\frac{\pi}{2} \pm n\pi\right)}{-\sin\left(\frac{\pi}{2} \pm n\pi\right)} = -1, \qquad n = 0,1,2, \ldots$$

. . .

In the case of poles of order $m > 1$, we need to take some derivatives to get a formula for the associated residue.

Theorem 147. If f has a pole of order $m > 1$ at z_0, then

$$\boldsymbol{Res(f; z_0) = \lim_{z \to z_0} \frac{1}{(m-1)!} \frac{d^{m-1}}{dz^{m-1}} [(z - z_0)^m f(z)]}$$

Proof: By definition of a pole of order m, f has a Laurent series about z_0 of the form

$$f(z) = \frac{c_{-m}}{(z - z_0)^m} + \cdots + \frac{c_{-2}}{(z - z_0)^2} + \frac{c_{-1}}{(z - z_0)} + c_0 + c_1(z - z_0) + c_2(z - z_0)^2 + \cdots$$

Multiplying the above by $(z - z)^m$, we get

$$(z - z_0)^m f(z) = c_{-m} + \cdots + c_{-2}(z - z_0)^{m-2} + c_{-1}(z - z_0)^{m-1} + c_0(z - z_0)^m + \cdots$$

Taking the derivative of the above $m - 1$ times, gives us

$$\frac{d^{m-1}}{dz^{m-1}} [(z - z_0)^m f(z)] = (m-1)!\, c_{-1} + m!\, c_0(z - z_0) + \frac{(m+1)!}{2!} c_1(z - z_0)^2 + \cdots$$

Taking the limit of the above as $z \to z_0$, we have our result, i.e.,

$$\lim_{z \to z_0} \frac{d^{m-1}}{dz^{m-1}} [(z - z_0)^m f(z)] = (m-1)!\, c_{-1}. \blacksquare$$

As an example, consider

$$f(z) = \frac{e^z}{z^4 (z - i)^2}$$

By Part ii of Theorem 146, $f(z)$ has a pole of order four at 0 and a pole of order two at i.

By Theorem 147, we have that

$$Res(f; 0) = \lim_{z \to 0} \frac{1}{3!} \frac{d^3}{dz^3} [z^4 f(z)] = \frac{1}{6} \lim_{z \to 0} \frac{d^3}{dz^3} \left[\frac{e^z}{(z - i)^2}\right]$$

$$= \frac{1}{6} \lim_{z \to 0} \left[\frac{e^z (z^3 - (6 + 3i)z^2 + (15 + 12i)z - 18 - 17i)}{(z - i)^5}\right] = \frac{1}{6} \left[\frac{-18 - 17i}{-i}\right] = \frac{17}{6} - 3i$$

$$Res(f; i) = \lim_{z \to i} \frac{1}{1!} \frac{d}{dz} [(z - i)^2 f(z)] = \lim_{z \to i} \frac{d}{dz} \left[\frac{e^z}{z^4}\right] = \lim_{z \to i} \left[\frac{e^z (z - 4)}{z^5}\right] = e^i (1 + 4i)$$

Taking the derivatives can be tedious and prone to error. However, one can also obtain the result using an online symbolic computation program such as Wolfram Alpha.

Problem: Solve for the residues of the following function:

$$f(z) = \frac{\sin z}{z^3(z-1)^2}$$

Answer: $Res(f;-1) = \frac{1}{4}(\cos(1) - \sin(1))$, $Res(f;1) = \frac{1}{4}(\cos(1) - \sin(1))$

. . .

It is also possible to compute the residue of an essential singularity. This is typically done by determining the function's Laurent series around the essential singularity, and then reading off the value of c_{-1}. For example, take the function

$$f(z) = z^3 e^{\left(\frac{1}{z^2}\right)}$$

The Laurent series for this function about the essential singularity $z = 0$ is

$$z^3\left[1 + \frac{1}{z^2} + \frac{1}{2!}\frac{1}{z^4} + \frac{1}{3!}\frac{1}{z^6} + \cdots\right] = z^3 + z + \frac{1}{2!}\frac{1}{z} + \frac{1}{3!}\frac{1}{z^3} + \cdots$$

So, $Res(f;0) = \frac{1}{2}$.

4.10.2 Using Residues to Compute the Integrals of Trigonometric Functions

The residues of complex functions can be used to determine the integrals of rational real functions comprised of trigonometric functions, i.e., integrals of the form

$$\int_0^{2\pi} f(\sin\theta, \cos\theta)\ d\theta$$

To solve the integral, we first convert the above integral into a complex contour integral over the circle $|z| = 1$ as follows.

Let $z = e^{i\theta}, 0 \le \theta \le 2\pi$ which implies $\frac{1}{z} = \frac{1}{e^{i\theta}} = e^{-i\theta}$.

Next, we write the trigonometric functions in terms of z, e.g.,

$$\cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2} = \frac{1}{2}\left(z + \frac{1}{z}\right)$$

$$\sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2} = \frac{1}{2i}\left(z - \frac{1}{z}\right)$$

Further, when integrating along the circular contour $|z| = 1$, we have that $\frac{dz}{d\theta} = ie^{i\theta} = iz$ which implies $d\theta = \frac{1}{iz}dz$.

Let's see how this works with an example. Consider the integral

$$\int_0^{2\pi} \frac{\sin\theta}{2+\cos\theta}$$

Making the substitutions described above, we have

$$\int_{|z|=1} \frac{\frac{1}{2i}\left(z-\frac{1}{z}\right)}{2+\frac{1}{2}\left(z+\frac{1}{z}\right)} \cdot \frac{1}{iz}\, dz = \int_{|z|=1} \frac{z^2-1}{z(z^2+4z+1)}\, dz$$

Of the three simple poles (i.e., $0, -2 \pm \sqrt{3}$) only 0 and $-2+\sqrt{3}$ fall within the contour $|z| = 1$. Thus, by Cauchy's residue theorem, we have

$$\int_0^{2\pi} \frac{\sin\theta}{2+\cos\theta} = \int_{|z|=1} \frac{z^2-1}{z(z^2+4z+1)}\, dz = Res(f;0) + Res\left(f;-2+\sqrt{3}\right) = -1+1 = 0$$

where

$$f(z) = \frac{z^2-1}{z(z^2+4z+1)}$$

$$Res(f;0) = \lim_{z\to 0} zf(z) = \lim_{z\to 0} \frac{z^2-1}{z^2+4z+1} = -1$$

$$Res\left(f;-2+\sqrt{3}\right) = \lim_{z\to -2+\sqrt{3}} \left(z+2-\sqrt{3}\right) f(z) = 1$$

Life is the art of drawing without an eraser. – John Gardner

Acronyms and Symbols

Study without desire spoils the memory, and it retains nothing that it takes in.
– Leonardo da Vinci

∀ - for every

∋ - such that

∃ - there exists

∄ - there does not exist

$A \in B$ – A is an element of set B

$A \notin B$ – A is not an element of set B

$A \subset B$ – A is a proper subset of B, i.e., A cannot equal B

$A \subseteq B$ – A is a subset of B and could possibly equal B

$X \backslash Y$ – the elements of set X with any elements in common with set Y removed

$\mathbb{C}$ - the set of complex numbers

$\mathbb{Q}$ - the set of rational numbers

$\mathbb{N}$ - the set of natural numbers, i.e., 1,2,3, …

$\mathbb{R}$ - the set of real numbers

$\mathbb{Z}$ - the set of integers, i.e., …-2,-1,0,1,2,...

AAS – Angle-Angle-Side

ASA – Angle-Side-Angle

SAS – Side-Angle-Side

SSS – Side-Side-Side

References

[1] Fratini, S., *Shape Up and Solve It!: Learn Geometry Through Puzzles,* self-published on Amazon, https://www.amazon.com/dp/B0CRS7DRWF, January 2024. Electronic version is available free of charge at https://www.artofmanagingthings.com/home/my-books.

[2] *Affine geometry*, Wikipedia, https://en.wikipedia.org/wiki/Affine_geometry, accessed on 10 January 2024.

[3] *Playfair's axiom*, Wikipedia, https://en.wikipedia.org/wiki/Playfair%27s_axiom, accessed on 10 January 2024.

[4] *Non-Euclidean geometry*, Wikipedia, https://en.wikipedia.org/wiki/Non-Euclidean_geometry, accessed on 10 January 2024.

[5] *Gaussian curvature*, Wikipedia, https://en.wikipedia.org/wiki/Gaussian_curvature, accessed on 7 February.

[6] *Parallel postulate*, Wikipedia, https://en.wikipedia.org/wiki/Parallel_postulate, accessed on 13 January 2024.

[7] *Law of cosines*, Wikipedia, https://en.wikipedia.org/wiki/Law_of_cosines, accessed on 13 January 2024.

[8] *Saccheri quadrilateral*, Wikipedia, https://en.wikipedia.org/wiki/Saccheri_quadrilateral, accessed on 13 January 2024.

[9] *Euclid's elements*, Wikipedia, https://en.wikipedia.org/wiki/Euclid's_Elements, accessed on 13 January 2024.

[10] Wolfe, H., *Introduction to Non-Euclidean Geometry**, The Dryden Press, 1945. https://archive.org/details/in.ernet.dli.2015.222414

[11] Gans, D., *An introduction to Non-Euclidean Geometry**, Academic Press, 1973. https://archive.org/details/introductiontono0000gans

[12] Casey, J., *Project Gutenberg's First Six Books of the Elements of Euclid*, updated 18 July 2022. https://gutenberg.org/files/21076/21076-pdf.pdf

[13] Euclid's Elements, hosted by Clark University, http://aleph0.clarku.edu/~djoyce/java/elements/toc.html, accessed on 22 February 2024.

[14] *Hilbert's axioms*, Wikipedia, https://en.wikipedia.org/wiki/Hilbert%27s_axioms, accessed on 8 February 2024.

[15] *Polygon: Angles*, Wikipedia, https://en.wikipedia.org/wiki/Polygon#Angles, accessed on 15 March 2024.

[16] *Polygon triangulation*, Wikipedia, https://en.wikipedia.org/wiki/Polygon_triangulation, accessed on 16 February 2024.

[17] *Cauchy's functional equation*, Wikipedia, https://en.wikipedia.org/wiki/Cauchy%27s_functional_equation, 21 February 2024.

[18] *Hyperbolic triangle*, Wikipedia, https://en.wikipedia.org/wiki/Hyperbolic_triangle, accessed on 2 March 2023.

[19] *Lobachevskii function*, The Encyclopedia of Mathematics, https://encyclopediaofmath.org/wiki/Lobachevskii_function, accessed on 4 March 2024.

[20] *Perpendicular Bisectors of Triangle Meet at Point*, Proof Wiki, https://proofwiki.org/wiki/Perpendicular_Bisectors_of_Triangle_Meet_at_Point, accessed 6 March 2024.

[21] Conversation with Google Gemini, response to question "What is the relationship between Gaussian curvature and hyperbolic geometry?", 6 March 2024.

[22] *Hyperbolic geometry: Circles and disks*, Wikipedia, https://en.wikipedia.org/wiki/Hyperbolic_geometry#Circles_and_disks, accessed on 6 March 2024.

[23] *Hyperbolic functions*, Wikipedia, https://en.wikipedia.org/wiki/Hyperbolic_functions, accessed on 6 March 2024.

[24] *Beltrami–Klein model*, Wikipedia, https://en.wikipedia.org/wiki/Beltrami%E2%80%93Klein_model, accessed on 7 March 2024.

[25] *Poincaré disk model*, Wikipedia, https://en.wikipedia.org/wiki/Poincar%C3%A9_disk_model, accessed on 7 March 2024.

[26] *Hyperbolic geometry: Models of the hyperbolic plane,* Wikipedia, https://en.wikipedia.org/wiki/Hyperbolic_geometry#Models_of_the_hyperbolic_plane, accessed on 7 March 2024.

[27] *Horocycle*, Wikipedia, https://en.wikipedia.org/wiki/Horocycle, accessed on 11 March 2024.

[28] Gans, D., *Axioms for Elliptic Geometry*, Canadian Journal of Mathematics , Volume 4 , 1952 , pp. 81 – 92. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/A2FC37629A1D95487F1EC5DF6985C953/S0008414X00036075a.pdf/axioms_for_elliptic_geometry.pdf

[29] Coxeter, H.S.M, *Introduction to Geometry**, John Wiley & Sons, 1969.

[30] *Triangle inequality*, Wikipedia, https://en.wikipedia.org/wiki/Triangle_inequality, accessed on 27 March 2024.

[31] *Spherical trigonometry*, Wikipedia, https://en.wikipedia.org/wiki/Spherical_trigonometry, accessed on 28 March 2024.

[32] *Law of sines: The spherical law of sine*s, Wikipedia, https://en.wikipedia.org/wiki/Law_of_sines#The_spherical_law_of_sines, accessed on 28 March 2024.

[33] *Spherical law of cosines*, Wikipedia, https://en.wikipedia.org/wiki/Spherical_law_of_cosines, accessed on 28 March 2024.

[34] *Intro to Topology - Turning a Mug Into a Doughnut*, YouTube video by Fireside, https://youtu.be/IxAwhW4gP_c?si=Rr1tLI0pFUwkouRC, accessed on 30 March 2024.

[35] *Topology*, Wikipedia, https://en.wikipedia.org/wiki/Topology, accessed on 30 March 2024.

[36] *Topology*, Wolfram MathWorld, https://mathworld.wolfram.com/Topology.html, accessed on 30 March 2024.

[37] Vasanthi, R., *Homeomorphisms Between Letters of Alphabet: Topological Invariants of Classification of Letters*, European Chemical Bulletin, Volume 12, Special Issue 8, 2203. https://www.eurchembull.com/issue-content/homeomorphisms-between-letters-of-alphabet-topological-invariants-of-classification-of-letters-8506

[38] Fratini, S., *Mathematical Vignettes II*, self-published on Amazon, https://www.amazon.com/dp/B0CM1HCQR7, October 2023. Electronic version is available free of charge at https://github.com/sfratini33/art-of-managing-things-external/blob/master/free_books/MathVig-II.pdf.

[39] *Surface (topology)*, HandWiki, https://handwiki.org/wiki/Surface_(topology), accessed on 29 April 2024.

[40] *Platonic solid*, Wikipedia, https://en.wikipedia.org/wiki/Platonic_solid, accessed 30 April 2024.

[41] *Polyhedron*, Wikipedia, https://en.wikipedia.org/wiki/Polyhedron, accessed on 1 May 2024.

[42] *Euler characteristic*, Wikipedia, https://en.wikipedia.org/wiki/Euler_characteristic, accessed 30 April 2024.

[43] Levin, O., *Discrete Mathematics: An Open Introduction*, 3rd edition, https://discrete.openmathbooks.org/dmoi3/, accessed on 2 May 2024.

[44] *Why can I remove an edge from a cycle that is part of a connected graph?*, StackExchange Mathematics, https://math.stackexchange.com/questions/3877565/why-can-i-remove-an-edge-from-a-cycle-that-is-part-of-a-connected-graph, accessed on 2 May 2024.

[45] Eppstein, D., *Twenty-one Proofs of Euler's Formula*, The Geometry Junkyard website, https://ics.uci.edu/~eppstein/junkyard/euler/, accessed on 3 May 2024.

[46] *Toroidal polyhedron*, Wikipedia, https://en.wikipedia.org/wiki/Toroidal_polyhedron, accessed on 3 May 2024.

[47] *Fundamental polygon*, Wikipedia, https://en.wikipedia.org/wiki/Fundamental_polygon, accessed on 12 May 2024.

[48] *Möbius strip*, Wikipedia, https://en.wikipedia.org/wiki/M%C3%B6bius_strip, accessed on 11 May 2024.

[49] *Klein bottle*, Wikipedia, https://en.wikipedia.org/wiki/Klein_bottle, accessed on 12 May 2024.

[50] *What does the 4D Klein Bottle look like?*, YouTube video by mtbdesignworks, https://youtu.be/N_4VaG7ZQE8?si=pMZr2RqN69xYV2Lh, accessed on 12 May 2024.

[51] *M435 Ep 3 of 8 The Projective Plane RP2 Topology*, YouTube video, originally produced by The Open University, https://youtu.be/dBH-Id8VC3U?si=l-6YlEDX7J9LrAYv, accessed on 13 May 2024.

[52] Adams, C., Franzosa, R., *Introduction to Topology Pure and Applied*, Pearson Prentice Hall, 2008.

[53] *Orientability*, Wikipedia, https://en.wikipedia.org/wiki/Orientability, accessed on 14 May 2024.

[54] *Surfaces*, online course from The Open University, https://www.open.edu/openlearn/science-maths-technology/mathematics-statistics/surfaces/content-section-0, accessed 13 May 2204.

[55] *Subdivision surface*, Wikipedia, https://en.wikipedia.org/wiki/Subdivision_surface, accessed on 16 May 2024.

[56] Hatcher, A., *Algebraic Topology*, Cambridge University Press, 2001, https://pi.math.cornell.edu/~hatcher/AT/AT+.pdf, accessed on 16 May 2024.

[57] Earl, R., *Topology: A Very Short Introduction*, Oxford University Press, 2019.

[58] Chen Hui, George Teo, unpublished paper from the University of Chicago, https://math.uchicago.edu/~may/VIGRE/VIGRE2011/REUPapers/Teo.pdf, accessed on 18 May 2024.

[59] *Chebyshev distance*, Wikipedia, https://en.wikipedia.org/wiki/Chebyshev_distance, accessed on 19 May 2024.

[60] *Taxicab geometry*, Wikipedia, https://en.wikipedia.org/wiki/Taxicab_geometry, accessed on 19 May 2024.

[61] *Hamming distance*, Wikipedia, https://en.wikipedia.org/wiki/Hamming_distance, accessed on 19 May 2024.

[62] *Infimum and supremum*, Wikipedia, https://en.wikipedia.org/wiki/Infimum_and_supremum, accessed on 19 May 2024.

[63] *Metric Spaces: A subset U of Y is open in Y if and only if U = V n Y for some open subset V of X*, YouTube video by The Math Sorcerer, https://youtu.be/mYWwJiJdqP0?si=nzr7_ijyQPXrw7sL, accessed on 22 May 2024.

[64] *The Closure of a Set is Closed: Metric Spaces Proof*, YouTube video by The Math Sorcerer, https://youtu.be/SziE_llSURA?si=dwtjn4u2Qy8rYxNB, accessed on 22 May 2024.

[65] *A Set is Closed if and only if its Complement is Open: Metric Spaces*, YouTube video by The Math Sorcerer, https://youtu.be/Yn9nJnv6OQU?si=mSVKY5BavfY1R6k-, accessed on 23 May 2024.

[66] *Cauchy sequence*, Wikipedia, https://en.wikipedia.org/wiki/Cauchy_sequence, accessed on 24 May 2024.

[67] *Euclidean Space is Complete Metric Space*, Proof Wiki, https://proofwiki.org/wiki/Euclidean_Space_is_Complete_Metric_Space, accessed on 26 May 2024.

[68] *Convergent Sequence in Metric Space has Unique Limit*, Proof Wiki, https://proofwiki.org/wiki/Convergent_Sequence_in_Metric_Space_has_Unique_Limit, accessed on 25 May 2024.

[69] *Baire category theorem*, Wikipedia, https://en.wikipedia.org/wiki/Baire_category_theorem, accessed on 25 May 2024.

[70] *L^p space*, Wikipedia, https://en.wikipedia.org/wiki/Lp_space, accessed on 26 May 2024.

[71] Gamelin, T.W., Greene, R.E., *Introduction to Topology**, 2nd Edition, Dover Publications, Inc., 1999. https://archive.org/details/introductiontoto00game

[72] *Compact space*, Wikipedia, https://en.wikipedia.org/wiki/Compact_space, accessed on 28 May 2024.

[73] *The Concept So Much of Modern Math is Built On: Compactness*, YouTube video by Morphocular, https://youtu.be/td7Nz9ATyWY?si=0jPb53HNuMtO89jJ, accessed on 28 May 2024.

[74] *Heine–Borel theorem*, Wikipedia, https://en.wikipedia.org/wiki/Heine%E2%80%93Borel_theorem, accessed on 29 May 2024.

[75] *Heine–Cantor theorem*, Wikipedia, https://en.wikipedia.org/wiki/Heine%E2%80%93Cantor_theorem, accessed on 31 May 2024.

[76] *Long line (topology)*, Wikipedia, https://en.wikipedia.org/wiki/Long_line_(topology), accessed on 1 June 2024.

[77] *Cocountable topology*, Wikipedia, https://en.wikipedia.org/wiki/Cocountable_topology, accessed on 5 June 2024.

[78] *Subspace topology*, Wikipedia, https://en.wikipedia.org/wiki/Subspace_topology, accessed on 7 June 2024.

[79] *Separation axiom*, Wikipedia, https://en.wikipedia.org/wiki/Separation_axiom, accessed on 12 June 2024.

[80] *Urysohn's lemma*, Wikipedia, https://en.wikipedia.org/wiki/Urysohn%27s_lemma, accessed on 13 June 2024.

[81] Munkres, J., *Topology**, Prentice Hall, 2014. https://math.ucr.edu/~res/math205B-2018/Munkres%20-%20Topology.pdf

[82] Zhang, E., answer to question "Prove that if a set is nowhere dense iff the complement of the closure of the set is dense" on Mathematics Stack Exchange, https://math.stackexchange.com/questions/1630165/prove-that-if-a-set-is-nowhere-dense-iff-the-complement-of-the-closure-of-the-se, accessed on 19 June 2024.

[83] Rodgers, N., *Learning to Reason: An Introduction to Logic, Sets, and Relations**, John Wiley and Sons, Inc., 2000.

[84] *Connected space*, Wikipedia, https://en.wikipedia.org/wiki/Connected_space, accessed on 23 June 2024.

[85] *Vector Addition*, YouTube video from rootmath, https://youtu.be/lulSApFPw1M?si=7NF_7VQoCg3Ho4K9, accessed on 26 June 2024.

[86] *Root of Unity*, Wikipedia, https://en.wikipedia.org/wiki/Root_of_unity, accessed on 29 June 2024.

[87] *Fundamental theorem of algebra*, Wikipedia, https://en.wikipedia.org/wiki/Fundamental_theorem_of_algebra, accessed on 29 June 2024.

[88] *Visualizing simple complex functions*, YouTube video from MathMajor, https://youtu.be/5PfvpWx2Rdk?si=yad_IjSvC10_m4RU, accessed on 2 July 2024.

[89] *What does a complex function look like? #SoME3*, YouTube video from mathematimpa, https://youtu.be/r1h3eNQ2YM0?si=6OlmMR9_oyghPtjR, accessed on 2 July 2024.

[90] *The 5 ways to visualize complex functions: Essence of complex analysis #3*, YouTube video from Mathemaniac, https://youtu.be/NtoIXhUgqSk?si=TWzxOM3LeoyaJmk9, accessed on 2 July 2024.

[91] Brown, J.W., Churchill, R.V., *Complex Variables and Applications**, 8th Edition, McGraw-Hill, 2009.

[92] *Riemann sphere*, Wikipedia, https://en.wikipedia.org/wiki/Riemann_sphere, accessed 5 July 2024.

[93] *Cauchy–Riemann equations*, Wikipedia, https://en.wikipedia.org/wiki/Cauchy%E2%80%93Riemann_equations, accessed on 8 July 2024.

[94] *Jordan curve theorem*, Wikipedia, https://en.wikipedia.org/wiki/Jordan_curve_theorem, accessed on 1 August 2024.

[95] Saff, E.B., Snider, A.D., *Fundamentals of complex analysis with applications to engineering and science (3rd Edition)**, Pearson, 2003.

[96] *Proof of Cauchy Riemann Equations in Polar Coordinates*, Stack Exchange: Mathematics, https://math.stackexchange.com/questions/205671/proof-of-cauchy-riemann-equations-in-polar-coordinates, accessed on 4 August 2024.

[97] *Riemann integral*, Wikipedia, https://en.wikipedia.org/wiki/Riemann_integral#Integrability, accessed on 24 July 2024.

[98] *Morera's theorem*, Wikipedia, https://en.wikipedia.org/wiki/Morera's_theorem, accessed on 13 August 2024.

[99] *Picard theorem*, Wikipedia, https://en.wikipedia.org/wiki/Picard_theorem, accessed on 17 August 2024.

* Indicates the book or article is available for borrowing from the Internet Archive at https://archive.org/. In some cases, only an earlier edition of a book will be available.

Index of Terms