PETER ULRICKSON

A *Brief* Quadrivium

PETER ULRICKSON

A BRIEF QUADRIVIUM

THE CATHOLIC EDUCATION PRESS

THE CATHOLIC EDUCATION PRESS

ISBN 9781949822328

Doce me justificationes tuas.

To my father and to his.

Contents

List of Figures

Preface

Order fills our world. We find it in things, and in words. Order is both a gift and the goal of our striving.

Wise people have found that four simple kinds of mathematical study can help us to grasp this cosmic order. In the pages that follow, you will take the path that our predecessors have set out before us.

To succeed in this course of study you must be willing to work with patience and discipline. Read the words carefully. See if you can make them your own. There is a plan for how to work through this book in another book called *Teaching the Quadrivium*. Your teacher should have it. If you are studying alone, you should get it and follow the plan that is laid out there.

Do not rush or skip ahead. Do the exercises. Just as you practice in order to become a good athlete or a good musician, so you must practice to become a good mathematician.

This book and this preface are brief. Our lives are, too. In the time that remains, let us seek the best things.

In festo Sancti Bonaventurae
Anno Domini nostri Iesu Christi 2022

Part I

Geometry

Funes ceciderunt mihi in praeclaris.

1 Instruments

1.1 Introduction

Get a compass and straightedge. The straightedge need not be ruled. You will use it to make straight lines but not to measure. Sharpen your pencil well. If you use a pen instead, use one with a fine point. These simple instruments will acquaint you with the fundamental objects of geometry.

Do not read the book without completing the exercises. The purpose of the book is to develop habits. Habits are formed through repeated activity.

The purpose of this first chapter is to develop good habits in the use of your geometrical instruments. The exercises you will complete are not difficult. Despite their simplicity, they are important. Take the time to perform them well.

1.2 Exercises

1.2.1 Directions

Here are fifteen exercises to perfect your use of the compass and straightedge. Complete each exercise six times in a day. Do this for one week.

There is a checklist after the exercise descriptions. Use it to record your progress.

If you have time remaining during the week devoted to these exercises, you can use it to explore the puzzles given at the end of this chapter.

1.2.2 Exercise Descriptions

Exercise 1

Make a point.

- How small can you make the mark while ensuring that it is visible?
- Strive for symmetry.

- Intend to make the mark at a specific location. Check whether you succeed.

Exercise 2

Draw a line using the straightedge.

- Make the line fine.
- Keep the thickness uniform.
- Be sure that the straightedge does not move and that your pencil remains in contact with the edge.

Exercise 3

Draw a point, then draw a line passing through it.

- Be sure that the point and the line satisfy the conditions given already.
- The line must pass through the point. This means that the thickness of your pencil must be taken into account when putting the straightedge in place at the point.
- The point should remain faintly visible after the line is drawn through it. If you make your points so fine that the lines obscure them, make them slightly larger.

Exercise 4

Draw two points, then draw a single line that passes through both of them.

- The points and the line should satisfy the conditions of the earlier exercises.
- It is possible that the line will not pass perfectly through the center of each drawn point. Seek to keep any deviation uniform, so that if the line is slightly to one side of the center of one point it is also similarly situated with respect to the center of the other point.

To arrive at this uniformity, you must be careful when you accommodate the pencil thickness as you set up the straightedge relative to the two points.

Exercise 5

Draw a point, then draw two lines each passing through that point.

- Continue to strive for accuracy and uniformity in producing lines and points.
- The two lines should almost obscure the point entirely.

- The overlap of the two lines should not extend beyond the point.

Exercise 6

Draw a line segment, then pick up your straightedge from the page. Then place the straightedge along the segment, and use it to draw an extension of the segment.

- Be careful that the extension of the segment is in line with the initial segment.
- Be sure that the extension and the original segment coincide at the endpoint of the segment.
- Let the initial segment and its extension have the same thickness.

Exercise 7

Draw two segments that do not intersect and are not parallel. Extend them until they intersect.

- Depending on the initial configuration it might be necessary to connect another piece of paper.
- At least once, try to do this when the point of intersection is at a greater distance from one of the segments than the length of the straightedge. This means that you will be required to extend the segment (as you did in the previous exercise) repeatedly.

Exercise 8

Use the compass to draw a circle.

Drawing circles well with a compass takes some practice. Keep pressure on the fixed point and let the drawing point remain free and light, barely touching the paper. Hold the compass hinge (where the two legs meet) in the fingertips of one hand. Turn this as you would turn a top. The motion is in the fingertips more than in the wrist. It can be helpful to lean the compass slightly in the direction of the rotation, so that you pull the compass around as you go.

- Keep the point of the compass fixed in place.
- The radius should remain the same.
- The circumference drawn should be fine and uniform. To do this, you must keep the pressure on the drawing end light.

Exercise 9

Draw a point, then draw a circle having this point as its center.

- Once the compass point has been placed carefully on the drawn point, the principles are as in the previous exercise.

Exercise 10

Draw a segment, then draw a circle having this segment as radius.

- Fix the compass point carefully on one endpoint of the segment, then open or close the compass so as to arrive at the correct radius.
- After having completed the circle, observe how successful you have been in maintaining the same radius throughout the circle.

Exercise 11

Draw two points, draw the segment between them, then draw two circles. Each circle should have the segment as radius and one of the points as center.

- Let the points (in which the segment terminates) remain slightly visible.
- See that the circles drawn are of the same size.
- The circles intersect in two points. Use these points of intersection as a way to check for any asymmetry in your drawing.

Exercise 12

Draw a point, draw a circle with this point as center, and draw a line through the center of the circle.

- Be sure that the line passes through the center of the circle.
- Draw the line long enough so that it intersects the circle twice.

The segment between the two points of intersection with the circle is called a *diameter.*

Exercise 13

Draw three points that do not lie in a line. Connect the points with segments to produce a triangle.

- The sides of the triangle should intersect exactly at the points originally drawn.

Exercise 14

Draw a circle, and mark three points on the circle. Connect these points in a triangle.

- The points are to lie on the circle.
- The lines are to intersect the circle exactly at the points marked.

Exercise 15

Draw a circle and draw a line through the center of the circle, extending it to intersect the circle in two points. Mark a third point on the circle, and produce a triangle from the three points.

- Only two sides of the triangle need to be drawn here, since the third is already given as the diameter.

1.2.3 *Checklist*

Record your progress in the exercises.

	Day 1	Day 2	Day 3	Day 4	Day 5
Exercise 1					
Exercise 2					
Exercise 3					
Exercise 4					
Exercise 5					
Exercise 6					
Exercise 7					
Exercise 8					
Exercise 9					
Exercise 10					
Exercise 11					
Exercise 12					
Exercise 13					
Exercise 14					
Exercise 15					

1.3 *Puzzles*

Consider the following puzzles. Avoid seeking out solutions from other sources. Simply enjoy exploring them on your own. You will learn some of the solutions later.

Use only a compass and straightedge. If your straightedge is ruled, avoid making measurements. All of these tasks can be accomplished with your instruments.

- Make an equilateral triangle. This is a triangle with three equal sides.
- Make a square.
- Draw a triangle and produce a circle inside the triangle that touches each side of the triangle in a single point.
- Draw a triangle and produce a circle that passes through each corner of the triangle.
- Draw a segment. Divide the segment into two pieces of exactly the same size.
- Draw a segment. Divide the segment into three pieces of exactly the same size.
- Draw a circle, and eliminate any mark suggesting the center of the circle. Find the center of the circle.
- Draw a circle and a point outside the circle. Draw a line that passes through the point and touches the circle in exactly one point.

Such a line is said to be *tangent* to the circle. The word "tangent" comes from a root meaning "touch." Such meaning is present in the English word "tangible."

2

Procedures

2.1 Introduction

The procedures given in this chapter require that you use your instruments skillfully, as you have practiced. At this time, we will simply accept the procedures and their results. As you practice them, you will acquire conviction that the objects you draw satisfy the stated properties. Later, we will reflect on these same procedures in a new way, investigating how we can be certain that they work as intended.

Take time to complete the procedures carefully. Use the same care you were encouraged to use in the initial exercises. There is satisfaction to be found in producing clear, beautiful drawings.

Once again, complete all the exercises, repeatedly, daily, for a week. A checklist is at the end of the chapter.

2.2 Exercises

Exercise 1

Task: Draw a segment. Produce an equilateral triangle having this segment as one of its sides.
Procedure: Let the two ends of the segment be called A and B. Draw a circle with its center at A and with segment AB as radius. Draw a second circle. The second circle has its center at B, with the segment AB again as radius. These two circles intersect in two points. Choose one of the points. Call it C. Produce the segments AC and BC.

Exercise 2

Task: Given a segment, find the midpoint of the segment.
Procedure: Set the compass point at one end of the segment, and open the compass so that the radius is larger than half of the segment.

You in fact produce something more. In addition to the midpoint itself, you obtain the perpendicular bisector of the segment.

Produce a circle with this radius. At the other endpoint of the segment produce another circle with this same radius. These two circles intersect in two points. Use the straightedge to draw the line connecting these two points. The point at which the given segment and the newly drawn line intersect is the midpoint of the segment.
Variation: Vary the amount by which you open the compass in order to create the two circles. In the special case that you open the compass to match the given segment exactly, you arrive at the construction of an equilateral triangle, the previous procedure.

Exercise 3

Task: Draw a point, and draw two rays emanating from this vertex point, thereby producing an angle. Produce the ray that bisects the angle.
Procedure: Set the point of the compass at the vertex of the angle, and produce an arc of a circle which intersects both rays. There are two points in which the arc intersects the rays. Connect these points in a segment. Produce the midpoint of this segment. Draw a ray from the angle vertex through the midpoint. This is the angle bisector.
Variations:

It is not necessary to produce the entire circle.

- After producing the segment that terminates in the two sides of the angle, it is not necessary to go through all the steps to find the midpoint. It is sufficient to produce arcs of equal radius which intersect in a single point. The ray from the angle vertex through this point is both the angle bisector and the perpendicular bisector of the segment.
- Be sure to experiment with angles that are small, angles that are approximately right, and angles that are larger than right angles.

Exercise 4

Task: Draw an angle, and draw a separate ray. Produce a copy of the given angle having the given ray as one of its sides.
Procedure: Let the vertex of the given angle be called A. With center A produce a circular arc intersecting the two sides of the angle in points B and C. Let D be the point from which the separate ray emanates. With center D make a circle with radius AB. This circle cuts the ray in a point E so that segment DE is the same as segment AB. Use the compass to measure the segment BC, and produce an arc with this radius centered at E. Let F be the point at which this arc and the circle (centered at D) intersect. Draw the ray DF. The angle determined by rays DE and DF is the angle that was sought.

Observe that there are two that ways the angle can be copied. It can be copied on either side of the given ray.

Exercise 5

Task: Draw a line and mark a point on the line. Produce a line perpendicular to the given line that passes through the given point.
Procedure: With the marked point as center, produce a circular arc intersecting the line in two points A and B. Open the compass wider, and then produce circular arcs with centers at A and B, having the same radius, which intersect on one side of the line. Draw the line passing through this point of intersection and the marked point.
Variation: Complete this task when the point marked is near the end of the line drawn, or even at the very end. In this case you will need to extend the line further as a preliminary step before making a circular arc.

You produce a segment AB and then produce its perpendicular bisector. By the construction of AB, the perpendicular bisector passes through the marked point.

Exercise 6

Task: Draw a line and draw a point not on the line. Produce a line perpendicular to the given line that passes through the given point.
Procedure: Place the point of the compass at the marked point, and open the compass so that the radius extends beyond the line. Draw a circular arc intersecting the line in two points A and B. With center at A produce a circular arc having radius larger than half the segment AB. With center B produce an arc with the same radius which intersects the first one. Draw the line from this point of intersection through the given point. This line is perpendicular to the given line and passes through the given point.
Challenging Variation: On a large piece of paper, complete this task when the distance from the point to the line exceeds the span of your compass. The complete procedure is up to you, but here is a hint. Pick a point on the line that you think is close to the intersection between the given line and the line to be produced. Produce a perpendicular to the given line through this point. Use your straightedge to extend this line towards the given point. It is unlikely that the line you produce will be incident with the point. It will, however, be close. Continue from here, using your ability to construct perpendiculars and your familiar sense of properties of parallel and perpendicular lines.

Exercise 7

Task: Draw a point and a line not passing through the point. Produce a line passing through the given point and parallel to the given line.
Procedure: Combine the previous two exercises. Let the point be P and the given line be ℓ. Produce through P a line m perpendicular to ℓ. Then produce the line n through P perpendicular to the line m. The line n is the line sought.

Exercise 8

Task: Draw a segment. Draw a square having this segment as one of its sides.
Procedure: Let the ends of the given segment be A and B. Produce the line perpendicular to this one at B (which requires extending the given segment AB). Produce a circular arc with center at B and radius AB intersecting this perpendicular line in a point that we will call C. Then BC is a second side of the square to be constructed. Continue in the same manner, producing at C the perpendicular line to BC at C. Mark the point D on the same side of BC as A which is at distance AB from C. This yields a third side of the square. Finally, draw the line connecting A and D.

Exercise 9

Task: Draw a circle and remove any mark indicating its center. Find the center of the circle.
Procedure: Choose two points A and B somewhat near each other on the circle. Produce the perpendicular bisector of the segment AB (which you do when you find the midpoint of the segment AB). Choose two other points C and D on the circle, and produce the perpendicular bisector of the segment CD. Extend the two perpendicular bisectors until they intersect. Their intersection is the center of the circle.
Variation: Instead of beginning with a complete circle, begin with only a portion of a circle. The same method will allow you to find the center. Once the center has been found, the given arc can be completed to a whole circle.

The segments AB and CD that you produce are called *chords* of the circle.

Exercise 10

Task: Draw a triangle. Produce a single circle passing through each vertex of the triangle.
Procedure: Let the vertices of the triangle be A, B, and C. Produce perpendicular bisectors of segments AB and BC, and extend them sufficiently far so that they intersect. Let the point of intersection be D. Produce the circle with center D and radius DA.
Variations: Be sure to practice this procedure with triangles containing a right angle and with triangles containing an angle larger than a right angle.

Exercise 11

Task: Draw a triangle. Produce a circle that touches each side of the triangle in a single point.
Procedure: Let the vertices of the triangle be A, B, and C. Produce rays bisecting the angles at A and at B. Extend these angle bisectors

to the point D in which they intersect. Produce the line through D perpendicular to line AB, and let E be the point in which this line intersects the line AB. Produce the circle with center D and radius DE.

Exercise 12

Task: Draw a circle, and draw a point outside the circle. Produce a line that is tangent to the circle and that passes through the point.
Procedure: Let the center of the circle be A, and let the point given be B. Draw the line AB and let the point in which it intersects the circle be called C. Produce the circle with center A and radius AB. Produce the line perpendicular to AB which passes through the point C. Let the intersection of this line with the larger circle through B be called D. Produce the line AD, and let F be the point in which it intersects the smaller circle. Draw the line through B and F.

If you are unfamiliar with the word "tangent," look back to the note at the end of Chapter 1.

2.3 *Checklist*

Practice all the procedures above three times each day for five days. Be sure that you vary the data with which you begin. Choose a variety of segment lengths and angles. Use your instruments with care and precision.

	Day 1	Day 2	Day 3	Day 4	Day 5
Exercise 1	☐☐☐	☐☐☐	☐☐☐	☐☐☐	☐☐☐
Exercise 2	☐☐☐	☐☐☐	☐☐☐	☐☐☐	☐☐☐
Exercise 3	☐☐☐	☐☐☐	☐☐☐	☐☐☐	☐☐☐
Exercise 4	☐☐☐	☐☐☐	☐☐☐	☐☐☐	☐☐☐
Exercise 5	☐☐☐	☐☐☐	☐☐☐	☐☐☐	☐☐☐
Exercise 6	☐☐☐	☐☐☐	☐☐☐	☐☐☐	☐☐☐
Exercise 7	☐☐☐	☐☐☐	☐☐☐	☐☐☐	☐☐☐
Exercise 8	☐☐☐	☐☐☐	☐☐☐	☐☐☐	☐☐☐
Exercise 9	☐☐☐	☐☐☐	☐☐☐	☐☐☐	☐☐☐
Exercise 10	☐☐☐	☐☐☐	☐☐☐	☐☐☐	☐☐☐
Exercise 11	☐☐☐	☐☐☐	☐☐☐	☐☐☐	☐☐☐
Exercise 12	☐☐☐	☐☐☐	☐☐☐	☐☐☐	☐☐☐

3
The Foundation of a Science

3.1 Terms

We come to know about mathematical objects by speaking together about them. To speak in an orderly way, we must grasp clearly the words that we use. The distinctive words of an area of knowledge are called *terms*.

It is possible that you will find this chapter somewhat confusing. Do not become discouraged. To orient yourself, you can look ahead to the section in which we define some geometrical terms. Spend a reasonable amount of time and then keep moving.

Some of our terms are accepted as basic or foundational. We do not give a mathematical account of what these terms mean. Instead, we accept that they are understood in the same way by all people. Such terms are known as undefined terms. Other terms can be defined by making use of these undefined terms. In this way we build up a mathematical vocabulary that we share with others, as we move together in speech from the most familiar things towards things that are less clear to us.

To say that these are undefined does not mean that it is unreasonable to speak about them and to offer a verbal account of their character. It simply means that speech about these terms belongs to a different field of study.

Here are some terms that we use that are undefined.

- point
- segment
- line
- ray
- angle
- curve
- lies on (as in, "the point P lies on the line ℓ")
- the same (as said of segments)
- the same (as said of angles)

A line will be thought of as something that can be extended indefinitely, but not necessarily as something which has been extended already in this way.

We will also use the word "congruent" in place of "the same" for both segments and angles.

The list above is not exhaustive. We will not always be strict in using only these terms. We might, for example, speak of a line "passing through" a point, which is to be understood as equivalent to saying that the point "lies on" the line.

3.1.1 Definitions

Here are a few simple definitions of terms used in geometry. The purpose of these examples is for you to see that we define new terms using things that are previously accepted. In the first example, those previously accepted things are the undefined terms. Later definitions can build upon defined terms as well.

Definition 1. *To say that two lines intersect means that there is a point that lies on both of the lines.*

This is not a surprising statement. The thing that you are to notice is that it carefully uses more elementary ways of speaking. The key parts are "point," "line," and "lies on," which are undefined terms of geometry. We also see the phrase "there is." This is general intelligible speech whose scope extends beyond the subject matter of geometry.

As we have noted, to define *intersect* requires only undefined terms. We now use this defined term to give a new definition.

Definition 2. *To say that two lines are parallel means that they do not intersect.*

By recalling the definition of *intersect*, you can see that to say lines are parallel means that there is not a point that lies on both of the lines.

Next comes an important definition about angles. This definition uses only undefined terms.

Definition 3. *Suppose that a line is given, along with a ray emanating from a point on the line. Such a configuration forms two angles, one on each side of the ray. To say that one of the angles formed is right means that it is the same as (or "congruent to") the other angle.*

Observe that a definition can involve statements that clarify the setting in which the term is used. The preliminary remarks in the preceding definition indicate that "right" is said (of an angle) in situations in which a line and a second angle are present.

Here are two other important definitions involving angles. They use comparison to right angles.

Definition 4. *An angle is said to be acute when it is less than a right angle, and an angle is said to be obtuse when it is greater than a right angle.*

You might wish to think carefully about the way that the words "between" and "same" could be used to give a fuller account of what it means for one angle to be greater than or less than another one.

Circles are fundamental to our study. We now define what is meant by this term.

Definition 5. *Given a curve and a point, we say that the curve is a circle with the point as its center when all segments from the point to the curve are congruent.*

Those equal segments emanating from the center get a special name, one that is likely familiar to you.

Definition 6. *A segment from the center of the circle to the circumference is called a radius of the circle.*

Definition 7. *A segment passing through the center of a circle and terminating in the circle on each end is said to be a diameter of the circle.*

You know how to construct an equilateral triangle using the compass and straightedge. We now specify exactly what we mean when we say that a triangle is equilateral.

Definition 8. *A triangle is said to be equilateral when each of its sides is the same as each of the other sides.*

In order to remember that "equilateral" refers to equality of sides, see that it contains the root "lateral" which means "side."

It might be that you have seen, in earlier study of geometry, that the angles in an equilateral triangle are also all the same. Observe that this is not part of our definition. Instead, as we will show, this property follows from the definition.

3.2 *Postulates*

Having agreed on the terms, the words that we will use, it is also important to set out clearly the conditions we understand to govern the terms. Here are five postulates for geometry.

Postulates are conditions that are set out in advance and used as principles for reasoning. A closely related term is "axiom." The word "axiom" is used more at the present time than the word "postulate."

The way that these postulates are phrased is somewhat unusual, but this wording is the standard one. They will make sense as we use them.

1. To draw a line from any point to any point.
2. To continue a segment in a line.
3. To draw a circle with a given segment as radius.
4. That all right angles are equal.
5. If a line intersects two lines, and the interior angles on one side are less than two right angles, then the lines, if produced, intersect on that side.

None of the postulates refers directly to our instruments. You can, nonetheless, see that some are related to the instruments you have used. With the exception of the fourth, each postulate refers to action in some way.

You might find the fourth postulate odd or seemingly unnecessary. Does this statement follow immediately from the definition of *right angle*? It does not. The definition of a right angle involves being congruent to (or "the same as") an angle that is adjacent to it. Right-ness, defined in this way, involves only equality with this specific neighboring angle. Think of two right angles far from each other. Each one

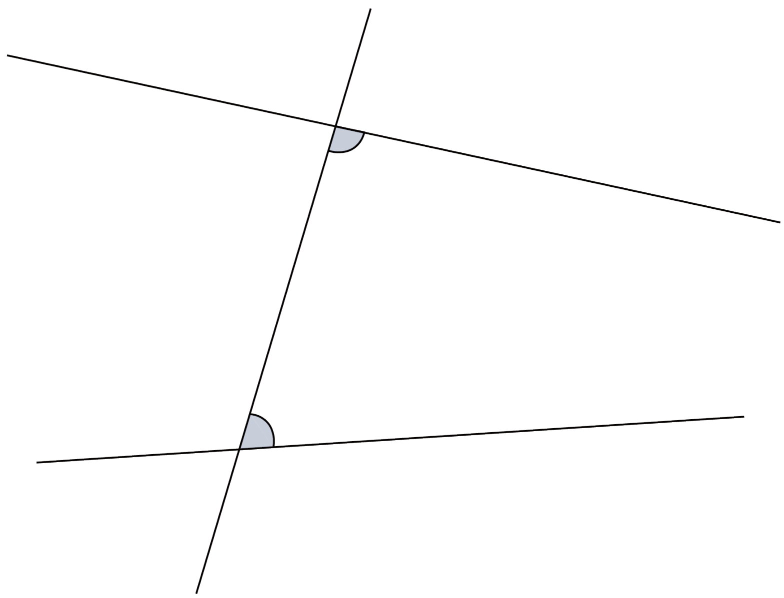


Figure 3.1: Postulate 5

is said to be right in virtue of how it compares to a nearby angle. We cannot conclude from this that the two distant angles are necessarily the same. Instead we must impose this as a condition for further mathematical speech. The condition corresponds to the sense of right angles that you obtained by performing constructions, and so we find it reasonable.

3.3 *Proof*

"What" and "Why" are two words that we use when asking questions. In your previous study of mathematics, you have been told to add, subtract, multiply, or divide. Such statements amount to the question "What is the number?"

In our study of geometry, our concern will not be to answer "What?" Instead, we will answer "Why?" We will answer the question: "Why does this statement follow from the definitions and postulates?" An answer to a question "Why?" comes as an explanation, not as a number or a symbol. It requires that we use words. We use the term "proof" to refer to the verbal explanations we give about why geometric statements follow from fixed principles, like our definitions and postulates.

Sometimes the word "demonstration" is also used in place of "proof."

How do we go about giving these explanations? How do we give a proof? You will come to understand this by seeing examples in the next chapter.

We use various words to refer to statements that have been proven. The main one we will use is "proposition." Other words that refer to proven statements are "theorem," "lemma," and "corollary." As you study mathematics further, you will learn the specific ways that these terms are used. For now, it suffices to realize that they all refer to mathematical statements that have a proof.

Sometimes people think that a mathematical statement follows from accepted principles, but they are unable to give a proof that this is so. In this case, if the statement is significant, it is called a conjecture. We will encounter conjectures in Part II, Arithmetic.

We must be confident that we, and those with whom we speak, have a sense of the forms that a reasonable explanation can take. The study of such forms is called logic. If you have the opportunity to study logic in the future, your current study of mathematical proof will be a good foundation.

3.4 Exercises

Exercise 1

Explain the difference between defined terms and undefined terms.

Exercise 2

List four undefined terms.

Exercise 3

Copy the five postulates.

Exercise 4

Give the five postulates from memory.

Exercise 5

List the names of statements that have proofs.

Exercise 6

Give the name of a mathematical statement that is thought to have a proof, but which has not yet been proven.

4 Proofs with Triangles

4.1 Origins and Conventions

A mathematician named Euclid wrote a book many years ago called *The Elements*. The title refers to the elements of geometry. This book has shaped the way that people talk and write about geometry and, more generally, all of mathematics. Our study of geometry follows the pattern of a small portion of Euclid's *Elements*.

You can hope to study all of Euclid's *Elements* in the future. The work you do now will prepare you to do so successfully.

Euclid's *Elements* is divided into thirteen parts, called "Books," each of which contains many propositions. To refer to a proposition from Euclid we use two numbers separated by a period. The first number, which is given as a Roman numeral, refers to the Book from *The Elements*. The second number refers to the number of the proposition within that book. As an example: Proposition II.11 refers to the eleventh proposition of the second book.

II is the Roman numeral for 2. You do not need to know many Roman numerals to understand our references.

Many propositions in this book are in Euclid. Some are not. The definitions, propositions, theorems, and lemmas in this book are all numbered sequentially, with a single series of numbers. This is to make them easy to find. When you see **Proposition 9** (I.1) it means two things. It is the ninth numbered item in this book, and it corresponds to Book I, Proposition 1 in Euclid's *Elements*.

You should try to remember the Euclidean numbering. While this book might be reorganized or rewritten someday, leading to a change in its own number system, the numbers that arise from Euclid's *Elements* will not ever change. That makes them the best for permanent references.

4.2 A First Proof

The first procedure you studied in Chapter 2 was to produce an equilateral triangle on a given segment. At that time we considered only a series of steps to take using physical instruments. We now work in a different way. Instead of making a thing, a drawing, we wish to see how properties of geometrical objects arise from the definitions and postulates that we have set out.

We will give the first proof in a highly structured way. Later we will prove things in a less formal way. The important thing to remember, the thing that unifies all the proofs, is that a proof is an account in words of how conclusions follow from principles. This means that a proof is not simply an abstract, logical entity. It is a form of persuasive writing, and therefore should adopt a style suited to the one who hears.

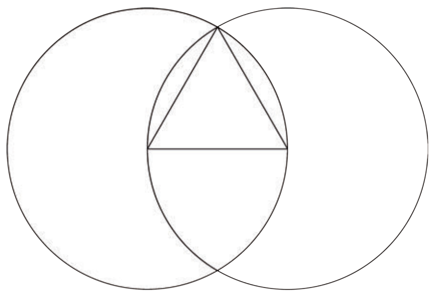


Figure 4.1: Proposition I.1

Proposition 9 (I.1). *On a given segment to construct an equilateral triangle.*

Proof. Consider the following claims in the left-hand column, with their corresponding justifications in the right hand column.

Claim	Justification
1. Let A and B be the endpoints of the segment.	Things given by hypothesis.
2. Draw the circle with center A and radius AB.	Postulate 3
3. Draw the circle with center B and radius AB.	Postulate 3
4. The two circles intersect in a point C.	Refer to figure.
5. Produce the segments AC and BC.	Postulate 1
6. AC is equal to AB, and BC is equal to AB.	Definition of circle.
7. AC is equal to BC.	Things each equal to a third thing are equal to each other.
8. The triangle ABC is equilateral, and on AB.	Definition of equilateral.

□

Observe that a symbol, a small square, is used to mark the conclusion of the proof. Some mathematical works also contain Q.E.D. or Q.E.F. as such markers. This helps the reader to recognize that all reasoning is now complete for the proof that was being given.

Q.E.D. and Q.E.F. abbreviate *Quod erat demonstrandum* and *Quod erat faciendum*, respectively. Translated, these mean "This is what was to be shown" or "This is what was to be made."

We said that A and B were given "by hypothesis." The hypotheses (plural) are the conditions that are assumed in a statement. A mathematical proposition is not simply a conclusion given absolutely, with unlimited scope. Instead, it is an assertion that the conclusion bears a relation to specifically stated conditions. These conditions are the hypotheses. In this proposition, the hypothesis was a segment, and a segment necessarily has two endpoints.

In the Part II: Arithmetic, we will see hypotheses like this: "Consider two numbers, one that is even and another that is odd."

4.2.1 *What is necessary in a proof?*

Let us consider another way of writing the proof. This second way is more succinct.

Proposition 10 (I.1). *On a given segment to construct an equilateral triangle.*

Proof. Let A and B be the endpoints of the segment. Draw the circle with center A and radius AB. Draw the circle with center B and radius BA. Let these circles intersect in C. Produce segments AC and BC. AC is equal to AB, and BC is equal to AB, so AC is equal to BC. Thus, the triangle ABC is equilateral and has AB as a side. □

The collection of claims made in this proof are the same as those made in the previous proof. The difference, which makes this second proof much shorter, is that the relationship between these claims and more elementary statements has been left implicit. This proof is neither better nor worse than the first one. It is simply different. It is a proof that relies on the reader's familiarity with the foundational elements we set forth previously.

A proof is an explanation, so it is an exchange of words. We offer a proof to some other person. At times this "other person" is you yourself, thought of as someone distinct. The best way to understand a proof is as a way to capture conversation between two people. Consider the example dialogue below. It is a conversation between me, teaching, and you, learning.

I	Let A and B be the endpoints of the segment. Draw the circle with center A and radius AB. Draw the circle with center B and radius BA. Let these circles intersect in C. Produce segments AC and BC. AC is equal to AB, and ...
You	Why are AC and AB equal?

I Each is a radius of the circle with center A and radius AB.

You Does this mean they are necessarily equal?

I Yes. The definition of a circle is that it is a curve with a central point such that all points from the center to the curve are equal. The point A is the center of the circle drawn, so all segments from A to the circle are the same.

You I see. Let's continue...

As you read proofs, you must seek to do so with an internal dialogue like the one above. Probe, inquire, and return to more foundational things. Over time you will find that you do not need to memorize proofs in their exact wording. Instead, you will simply remember what you have seen. When you see the meaning in the things, the words of the proof will come to you freely.

4.3 Accepted Propositions

We will not give proofs of Euclid's second, third, and fourth propositions. Instead we will simply accept them, and then use them in later proofs. If you study Euclid in greater detail later, you will learn about these proofs.

We omit not only the proof, but even the statement of Euclid's Proposition I.2. This proposition and its proof are elegant and remarkable, and are certainly worth your attention in the future. Euclid uses Proposition I.2 to prove the next proposition.

Proposition 11 (I.3)**.** *Given two unequal segments, to cut off from the greater a segment equal to the less.*

What this means, roughly, is that we can proceed as we did when working with a physical compass. Given a segment, we can measure it with the compass, pick up the compass, and make an arc with that radius at a different point on the piece of paper.

If you review the third postulate carefully you will see that it involves a specific segment as radius, rather than saying that the radius is equal to a given segment. This is a subtle point, so do not worry if you do not understand now.

Proposition 12 (I.4)**.** *If a side of one triangle is equal to a side of another triangle, and a second side of the first triangle is equal to a second side of the second triangle, and the angles included by these sides are equal, then the triangles are equal.*

To say that the triangles are equal means that all the sides and angles are equal, when considered in order. Because Proposition I.4 refers to a side, an included angle, and another side, it is sometimes referred to with the acronym SAS, for side-angle-side. It can then be called the SAS Congruence Principle.

4.4 *Isosceles Triangles*

To speak about isosceles triangles, we must give the definition of the term. This will ensure the agreement necessary for further discussion.

Definition 13. *A triangle is said to be isosceles when two and only two of its sides are equal.*

Some authors define isosceles to mean a figure in which two sides are equal, without requiring that only two be so. Under such a definition, equilateral triangles are called isosceles as well. This does not lead to any significant conflict. It is simply a case of a different convention. The things proved in each case will, suitably rephrased, be the same. Your teacher might wish to follow the other convention.

Observe that we exclude equilateral triangles from the isosceles triangles with the assertion that "only two" of the sides are equal. A triangle in which all three sides are distinct is called *scalene*. We will have little use for this term.

Perhaps you are aware from earlier study that the base angles of an isosceles triangle are equal to each other. This is not part of the definition of isosceles triangle. It is instead a property that follows from the things we have set down. As you read the following proof, remember to think of it as the outline of a possible conversation.

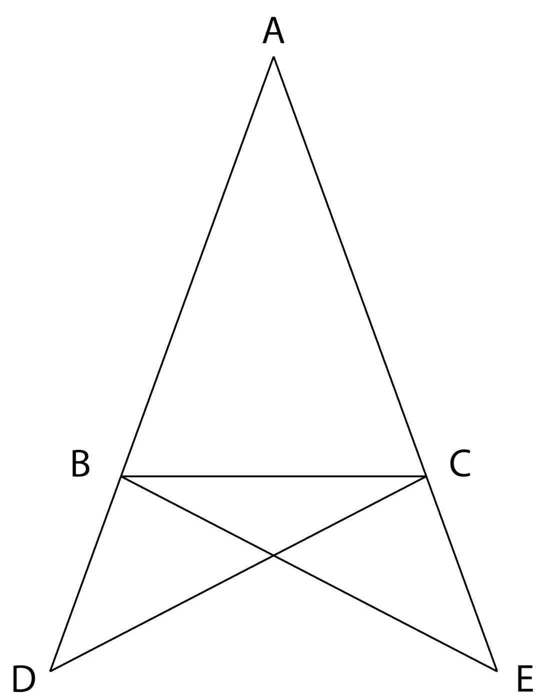


Figure 4.2: Proposition I.5

Proposition 14 (I.5). *In an isosceles triangle, the angles at the base are equal.*

In fact, Euclid's I.5 goes slightly beyond the statement here. The extension is not necessary for us.

Proof. Let the vertices of the triangle be A, B, and C, so that AB is equal to AC. Extend segment AB beyond B, and choose a point D beyond B on the line AB. Extend AC past C, and let E be the point beyond C on the line AC such that CE is the same as DB, by Proposition 11 (I.3). This is depicted in Figure 4.2.

The segment AD is composed of AB and BD. The segment AE is composed of AC and CE. Since AB is the same as AC, and BD is the same as CE, we see that AD is the same as AE.

Consider the triangles ABE and ACD. We know, from the preceding paragraph, that AD and AE are the same. We know by hypothesis that AC and AB are the same. Finally, the angle with vertex at A is the same as itself. We then conclude, by Proposition 12 (I.4), that triangles ABE and ACD are equal.

Since the triangles ABE and ACD are equal, the angle at D is equal to the angle at E, and the segment CD is equal to the segment BE. Since BD and CE are equal, we conclude by Proposition 12 (I.4) that triangles BDC and CEB are equal.

Earlier we saw that ABE and ACD are equal, so that the angle ABE and the angle ACD are equal. We have just seen that the trianlges BDC and CEB are equal, so that the angles CBD and BCE are equal. The whole angle ABE is composed of angle ABC and angle CBE. The whole angle ACD is composed of angle ACB and angle BCD. Taking parts that are the same away from wholes that are the same, we conclude that angle ABC is equal to angle ACB. □

When we name an angle with three points, the middle point is the vertex. So the angle ABE is the angle formed by two rays emanating from B, one of which passes through A and the other of which passes through E.

The definition of isosceles triangle involves equality of sides. From it we conclude equality of base angles. The next proposition shows us that from the equality of the base angles we can conclude equality of sides.

In the proof of the next proposition, we use a method that deserves mention. This method is known as "proof by contradiction" or *reductio ad absurdum*. This means that when we want to show that something follows from our principles, we assume that the opposing statement (known as the "negation") does follow from them, and see that a contradiction results. Since we are confident in the intelligibility of our principles, we conclude that the thing we assumed is to be rejected. It is easiest to understand this through examples, like the one coming now.

Proposition 15 (I.6). *If in a triangle two angles are equal, the sides opposite the equal angles are also equal.*

Proof. Let ABC be a triangle with equal angles at B and C.

(This is something we wish to reject.) Suppose that side AB is less than side AC.

Let D be the point on AC such that CD is equal to AB. By Proposition 12 (I.4), the triangle DCB is the same as the triangle ABC, since the sides DC and AB are the same, the side BC is the same as itself, and the included angles DCB and ABC are the same by hypothesis. Then the angle ABC is the same as the angle DBC, since each is the same as the angle DCB. This is impossible, since DBC is a part of ABC.

Therefore, AB is not less than AC.

We conclude the line of reasoning that began with the statement we wanted to reject.

Suppose, on the other hand, that AC is less than AB.

This is the other case to consider.

Reasoning as before, we would conclude that a part of the angle at C was equal to the whole angle ACB, which is impossible. Therefore, AC is not less than AB.

Since AB is not less than AC, and AC is not less than AB, they are equal.

$\square$

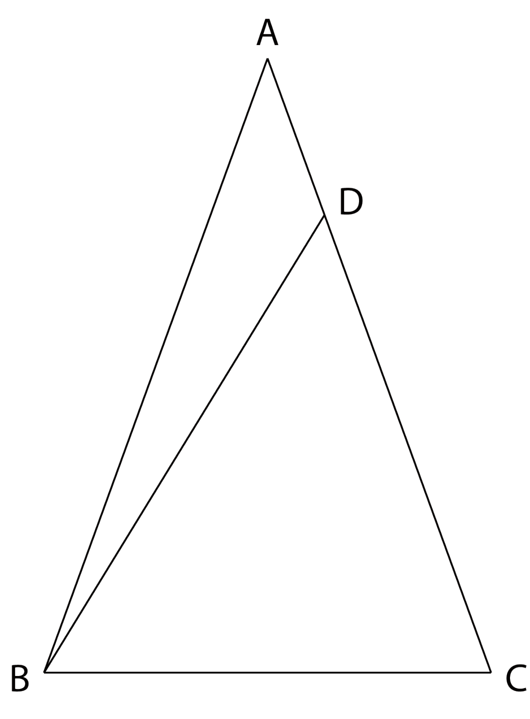


Figure 4.3: Proposition I.6

The preceding proof treated separate possibilities. Supposing that AB and AC are not equal, we considered first one possibility (AB is less than AC) and then the other. These two cases do not arise from anything essential in the geometric objects. Instead, they arise simply because of the freedom we have to give names to points. In order to keep proofs shorter and clearer we use the phrase "without loss of generality" when we can accommodate multiple cases by choosing names suitably. Here is that proof rewritten using this phrase.

Proof. Let ABC be a triangle with equal angles at B and C.

Suppose without loss of generality that side AB is less than side AC. Let D be the point on AC such that CD is equal to AB. The triangle DBC has equal sides DB and DC, thus the angle at B is equal to the angle BCD by Proposition I.5. Then then angle at B is equal to both angle BCD and angle BCE, so that the angles BCD and BCE are equal to each other. This is impossible, since BCD is a part of BCE.

Therefore, AB, taken to be the lesser side, is not less than AC. Therefore, the sides AB and AC are equal. □

If you wish to use the phrase "without loss of generality" when giving proofs, you must take care that in fact no generality has been lost. Here is one way of proceeding. Write out the proof in great detail, being very clear about all specifics. If you see that there is a significant portion of the argument that is repeated almost exactly, save for the names of the things being discussed, then try to find where a choice of names has led to this redundancy in the explanation.

Do not worry if the phrase "without loss of generality" is confusing to you at this time.

4.5 *Some Additional Results*

Consider a fixed segment with endpoints A and B, and let a triangle CDE be given. Suppose further that the segment AB is equal to the segment CD. Can we set up a triangle on the segment AB that is equal to the triangle CDE? By considering a diagram, you can convince yourself of this. There is more than one way to do it, unless we make further specifications. If we require that the new triangle be placed on a specific side of AB, and we also require that the angle at C be placed at A, then there is only one way to make the triangle. This is recorded in the following proposition.

Proposition 16 (I.7)**.** *There is a unique way to establish, on a given side of a given segment, a triangle with a side congruent to that segment, with an angle at a specified position.*

This is something that must be proved in order to give a complete account. We will not prove it. You are capable, though, of understanding the proof.

Something that follows from I.7 is the Side-Side-Side (SSS) congruence principle for triangles.

Proposition 17 (I.8)**.** *If the sides of one triangle are equal to the sides of another triangle, the triangles are equal.*

Again, we will omit the proof of this statement, but will be free to use it in further reasoning.

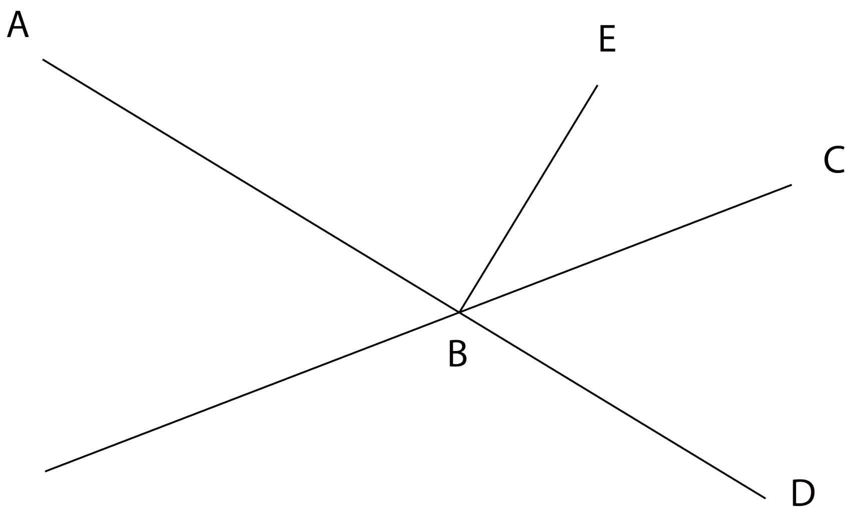


Figure 4.4: Proposition I.13

The situation of the following proposition is depicted in Figure 4.4.

Proposition 18 (I.13). *When one line intersects another, the angles produced are equal to two right angles.*

Proof. Let lines AB and CB intersect at B. Let D be a point on AB so that B is between A and D.

Either angle ABC and angle CBD are equal, or they are not equal. If the two angles are equal, we have shown what was required, since they are right.

If the two angles are not equal, let line BE be the line through B perpendicular to AB. The point C lies on one side of line BE. Let it be, without loss of generality, the opposite side as A. The angle ABE is right. Moreover, the angles EBC and CBD are together a right angle.

You will consider this statement on the production of perpendiculars, Proposition I.11, as an exercise.

In either case, then, the combination of ABC and CBD is equal to two right angles.

□

When two lines intersect, there are four angles at the intersection. A pair of those angles which do not share a side are called "vertical angles." As the next proposition will show, such angles are equal.

Proposition 19 (I.15). *Vertical angles are equal.*

Proof. Let lines AB and CB intersect at B, as in Figure 4.5, let D be beyond B on AB, and let E be beyond B on CB.

By Proposition 18 (I.13), angles ABC and CBD are together two right angles. Similarly, angles CBD and DBE are two right angles. Removing the common angle CBD, we conclude that angles ABC and DBE are equal. □

Subtracting the same thing from equal things yields equal remainders.

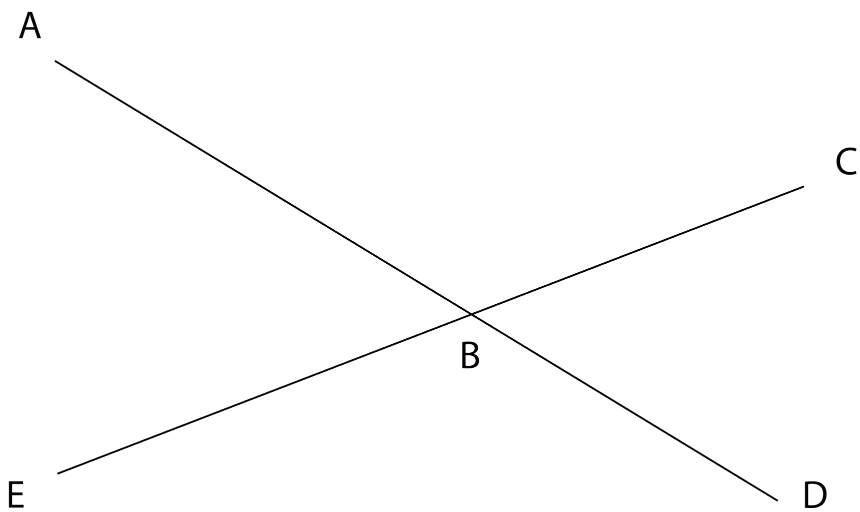


Figure 4.5: Proposition I.15

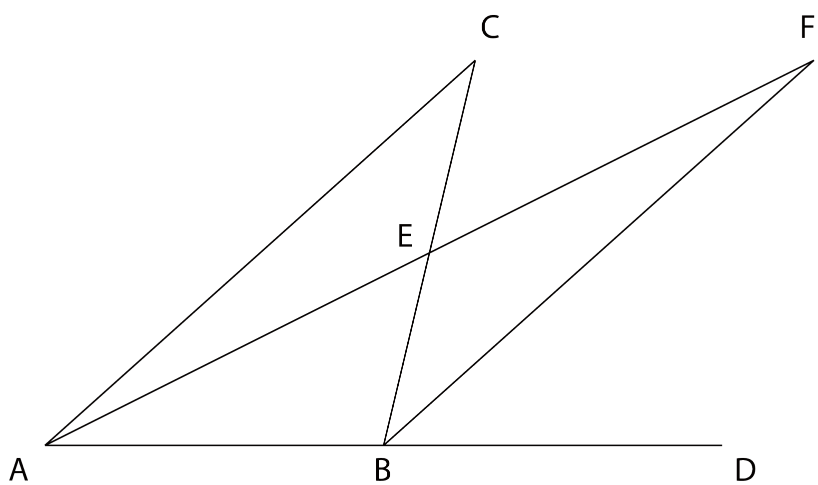


Figure 4.6: Proposition I.16

Proposition 20 (I.16)**.** *An exterior angle of a triangle is greater than either of the opposite interior angles.*

Proof. Let ABC be a triangle, and let side AB be continued to D. We wish to show that angle CBD is greater than angle BCA and angle BAC. We first consider angle BCA.

Let point E be the midpoint of the segment BC.

This midpoint statement depends on I.10, the proof of which is sketched in an exercise. For now you may content yourself with considering the diagram.

Extend AE beyond E to point F such that AE and EF are equal. The vertical angles CEA and BEF are equal by Proposition 19 (I.15). Since E is the midpoint of BC, and E the midpoint of AF, we conclude by Proposition 12 (I.4) that the triangles CEA and BEF are equal. This means in particular that angle ECA and angle EBF are equal.

Angle EBF is a part of angle EBD. Thus, angle EBD is greater than EBF and so is greater than ECA. This is what was to be shown.

To show that angle CBD is greater than angle CAB, extend CB beyond B to point G. Reasoning as above, we conclude that angle ABG is greater than CAB. Since ABG and CBD are vertical angles, hence equal, the proof is concluded.

□

4.6 *Exercises*

Exercise 1

Copy the first proof of Proposition 9 (I.1).

Exercise 2

Copy the second proof of Proposition 9 (I.1).

Exercise 3

Write a proof of Proposition 9 (I.1) from memory. Compare your proof to those given.

Exercise 4

With a partner, imitate the dialogue version of the proof of Proposition 9 (I.1). With each question asked, you should be able to justify the statement you have made by referring to a postulate or to a definition.

Exercise 5

Create your own dialogue version of a proof given in this chapter.

Exercise 6

Let two triangles be given. Let one triangle be ABC and let the other be DEF. Suppose the following conditions hold.

- The segments AB and DE are the same.
- The segments BC and DF are the same.
- The angle ABC and the angle FDE are the same.

Use Proposition I.4 to show that the two triangles are the same. In doing so, state clearly which vertex of the first triangle corresponds with which vertex of the second triangle.

Exercise 7

Proposition 14 (I.5) refers to isosceles triangles. Our definition of "isosceles" excludes equilateral triangles. Give a suitable statement about angles in equilateral triangles, and prove it by the same means as I.5.

Exercise 8

Complete this proof of Euclid's Proposition I.9, to produce a bisector of an angle.

Proof. Let the angle ABC be given, with vertex B. Without loss of generality let segments AB and BC be the same. On the segment AC construct the equilateral triangle ACD, with D on the opposite side of AC as B. *(Use Proposition 17 (I.8) to show that triangles ABD and CBD are the same, and thus the relevant angles are also the same.)* □

Think of A and C as arising from making a circle with center at B and intersecting with the two rays.

Exercise 9

The preceding exercise, in which you prove Proposition I.9, follows Euclid's approach. Observe that the proof there is slightly different from the procedure you studied earlier for producing angle bisectors with compass and straightedge. Justify your general procedure, which did not use an equilateral triangle, in the following way. Note that you must understand what statement you are proving.

Proof. Let an angle ABC be given, with AB and BC the same, and let a triangle ACD be given, with AD the same as CD ... □

Exercise 10

Complete the proof of Proposition I.10, to produce the midpoint of a segment.

Proof. Given a segment AB, produce on AB the equilateral triangle ABC. Produce the bisector of angle ACB, and let it intersect AB at D ... *(Argue that the triangles ACD and BCD are equal, and that the corresponding sides AD and BD are thus also equal.)* □

Exercise 11

The preceding exercise, in which you prove Proposition I.10, follows Euclid's approach. Observe that the proof there is different from the procedure you studied earlier for producing midpoints with compass and straightedge. Justify the general procedure in the following way.

Proof. Let a segment AB be given. Let a point C near B be chosen, and produce the circle with center A and radius AC. Let the point D on AB be such that BD is the same as AC, and produce the circle with center B and radius BD. Let the two circles intersect in points E and F. Let EF intersect AB at G... *(Use I.5 to show that angles EAB and EBA are the same. Use I.4 to show that angles AEG and BEG are the same. Conclude using I.4.)* □

Exercise 12

Given a point on a line, produce a perpendicular to the line through the point (Proposition I.11). Complete the proof.

Proof. Let A be a point on the given line, and let B be some other point on the line. Produce the circle with center A and radius AB, and let it intersect the line also in the point C. On segment BC produce the equilateral triangle BCD... *(Argue that triangles BDA and CDA are equal by Proposition 17 (I.8). Use the definition of right angle, and the fact that points A, B, and C are all in a single line.)* □

Exercise 13

Given a line and a point not on the line, produce a line perpendicular to the given line and passing through the given point (Proposition I.12). Complete the proof.

Proof. Let a point A be given, and a line. Choose a point B on the opposite side of the line from A, and produce the circle with center A and radius AB. Let it intersect the given line in points C and D. Let E be the midpoint of segment CD... *(Consider the triangles ACE and ADE. They are equal. Use the definition of right, keeping in mind that points C, E, and D are all in the same line.)* □

Exercise 14

Give diagrams that show the importance of fixing a side, and fixing the angle location, in order to obtain the uniqueness asserted in Proposition 16 (I.7).

Exercise 15

Explain why there is no loss of generality in the proof of Proposition 18 (I.13).

Exercise 16

Give the proof of Proposition 20 (I.16) from memory. Check your proof against the one given.

Exercise 17

Use Proposition 20 (I.16) to show that in a right triangle the other two angles are acute. *(Observe that such a triangle has an exterior angle that is right.)*

Exercise 18

Give a diagram showing that the point D in the proof of Proposition I.9 can be "outside" the angle ABC, if we do not specify that it be on the opposite side of the line AC. For which sorts of angles is this possibility relevant?

See Exercise 8.

Exercise 19

Make a diagram including the point G indicated in the proof of Proposition 20 (I.16), and make the rest of the proof explicit, following the pattern of reasoning already given.

5
Parallels

5.1 Properties of Parallels

We will now consider parallel lines. Recall that to say lines are parallel is to say that they do not intersect.

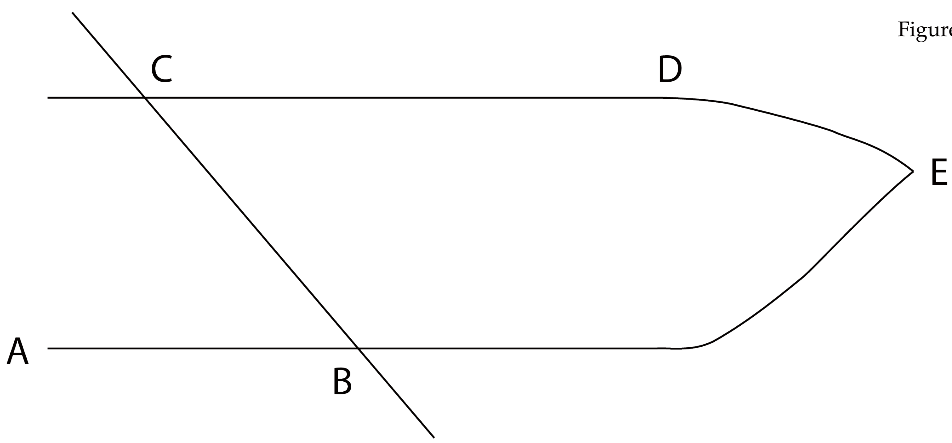


Figure 5.1: Proposition I.27

Proposition 21 (I.27). *Let a line intersect two others so that the alternate interior angles are equal. Then the lines are parallel.*

Proof. Let line BC intersect lines AB and CD in such a way that angles ABC and BCD are equal.

(We will show this leads to a contradiction.) Suppose that lines AB and CD intersect, and let the point of intersection be E.

Observe that the proof does not depend on the side of the line on which the (presumed) intersection occurs.

We then have the triangle BCE. The angle ABC is an exterior angle of this triangle, but it is equal by hypothesis to the opposite interior angle BCD. This contradicts Proposition 20 (I.16).

Therefore, the lines AB and CD do not intersect. □

Until this point we have not made use of the fifth postulate. Return to the previous chapter and read that postulate before continuing. The following proposition marks the first use of this postulate in Euclid's *Elements*.

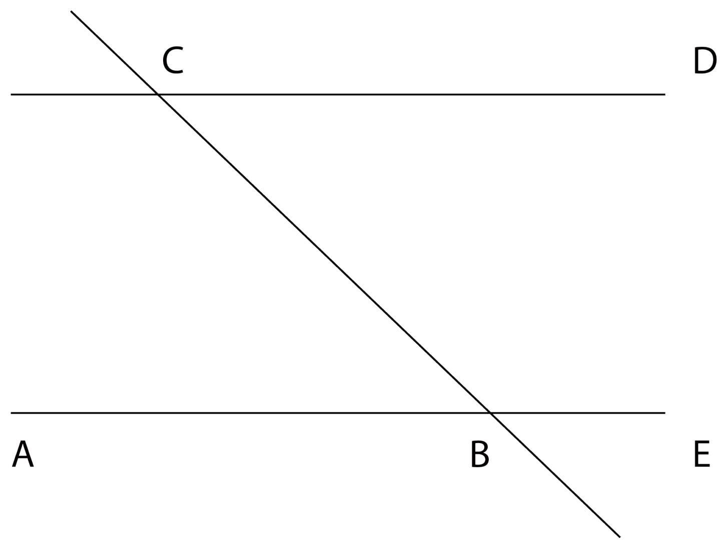


Figure 5.2: Proposition I.29

Proposition 22 (I.29). *A line intersecting two parallel lines makes the alternate interior angles equal.*

Observe that the lines are assumed to be parallel. The thing we are to show is that the condition on the angles follows from this. Be sure to distinguish this proposition from the previous one.

Proof. Let one of the lines be AB and the other CD, so that the tranverse line is BC intersecting them at points B and C.

(We wish to reject this.) Suppose that the angles ABC and BCD are not equal.

Then one of the angles is greater. Without loss of generality let ABC be the greater.

Let E be a point on AB beyond B. Since ABC and CBE are along the line AE, they are equal to two right angles, by Proposition 18 (I.13). Angle ABC is assumed greater than angle BCD, so that the angles CBE and BCD on the same side of the transversal BC are less than two right angles. According to Postulate 5, the lines AB and CD intersect on that side (the side of the point E) of the line BC. This contradicts the hypothesis that lines AB and CD are parallel.

Therefore, the angles ABC and BCD are equal.

□

Propositions I.27 and I.29 can be combined to say that two lines are parallel if and only if they yield equal alternate interior angles when cut by a transversal.

The next proposition shows that the relation "to be parallel to" satisfies a special property. In order to understand this special property, it is helpful to consider a non-mathematical example. One way that two people can be related is by ancestry. Another is by relationship within a generation, as cousins. By cousin let us mean, specifically, people having parents who are siblings.

Consider three people, persons A, B, and C. Suppose that person A is descended from person B, and person B is descended from person C. Can we conclude that person A is descended from person C? Yes, we can. Consider, on the other hand, persons D, E, and F. Suppose that person D is the cousin of person E, and person E is the cousin of person F. Can we conclude that person D is the cousin of person F? No, we cannot. If you have cousins, think about all their cousins in order to convince yourself of this.

When a relation between entities can be inferred based on their separate relations to some intermediate, we say that the relation is *transitive*. One way to phrase the following proposition, then, is to say that parallelism is a transitive relation among lines. It is thus more like "is descended from" than "is a cousin of."

The term *transitive* used here is modern. It is not a Euclidean term.

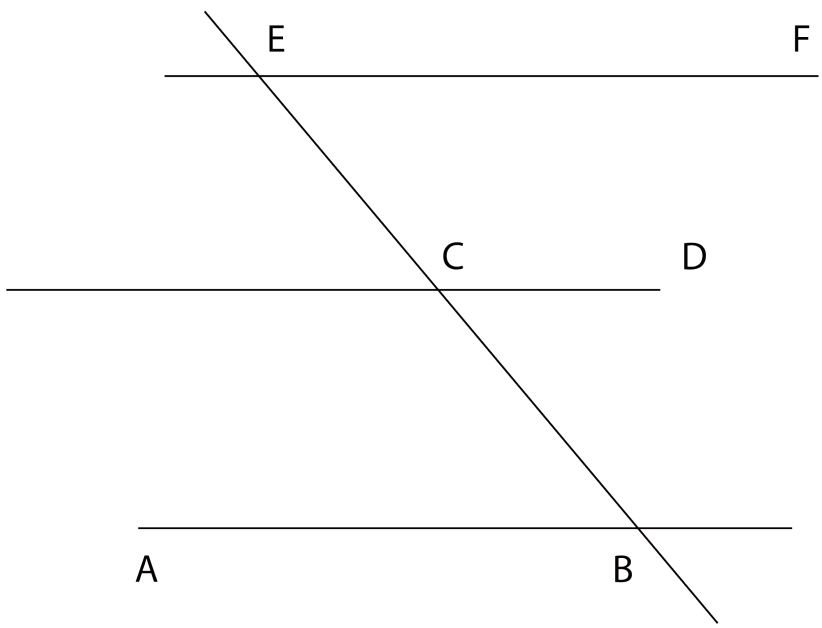


Figure 5.3: Proposition I.30

Proposition 23 (I.30)**.** *If a first line is parallel to a second line, and the second line is parallel to a third line, then the first line is also parallel to the third line.*

Proof. Let AB and CD be the first two lines, and consider the line BC intersecting the third line EF in the point E. By Proposition 22 (I.29) the alternate interior angles of the first two lines at B and C are equal. Similarly, via the equality of vertical angles given in Proposition 19, along with I.29, we see that the alternate interior angles of the second and third lines at C and E are equal. Thus, the alternate interior angles of the first and third lines at A and E are equal, so that the lines are parallel by Proposition 21 (I.27). □

It is possible to show that any two angles in a triangle must be less than two right angles. This is Euclid's Proposition I.17. To give the proof, we need only Proposition 20 (I.16), which relates interior and exterior angles. With the additional demonstrative power afforded by the fifth postulate and its consequences, we are now able to prove something stronger than I.17. Not only are two angles in a triangle less than two right angles; all three angles in a triangle are exactly equal to two right angles.

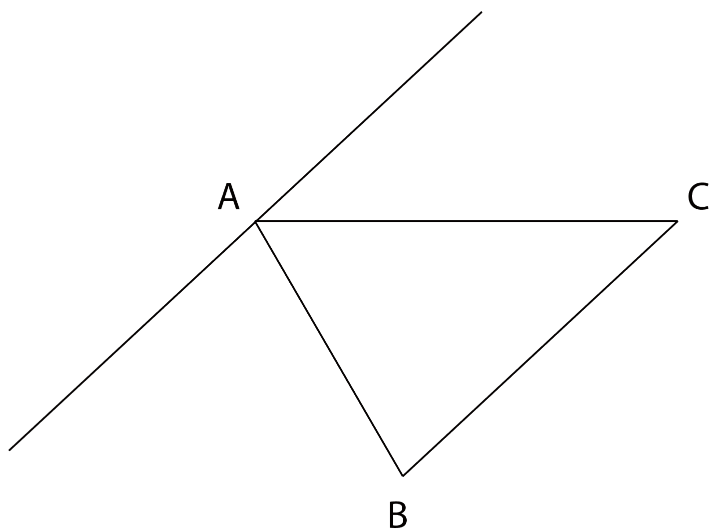


Figure 5.4: Proposition I.32

Proposition 24 (I.32)**.** *The angles in a triangle are two right angles.*

Proof. Let ABC be a triangle, and through A consider the line parallel to BC. By Proposition 18 (I.13) we know that the angles at A combine to two right angles. Applying Proposition 22 (I.29) to the lines AB and AC which are transverse to the two parallels, we see that the angles at A are equal to the alternate interior angles, which are equal to the other angles in the triangle. □

This proof requires the construction of a line parallel to a given line that passes through a given point. This is Proposition I.31, whose proof is Exercise 1 at the end of this chapter.

By I.32 we can discover something about triangles in a circle. It is thought that someone named Thales first discovered this result.

Proposition 25 (III.31)**.** *A triangle, one of whose sides is the diameter of a circle and whose other vertex lies on the circle, is right.*

This proposition is labelled with a Roman numeral III because it comes from the third book of Euclid's *Elements*.

Proof. Let triangle ABC be in a circle with AB the diameter. Let point D be the center of the circle. The triangles ADC and BDC are isosceles, since each has radii as two sides. Thus, angle ACD is the same as DAC, and angle BCD is the same as CBD. The collection of all the angles in the triangle ABC, then, is twice the combination of the angles ACD and BCD. By Proposition 24 (I.32), this is two right angles. Thus, the combination of the angle ACD and BCD is right, which was to be shown. □

5.2 *Exercises*

Exercise 1

Prove Proposition I.31, the construction of a parallel through a point.

Proof. Let a point A be given, and a line not through A. Choose a point B on the line, and produce the line AB. Consider some other point C on the line... *(At the point A, copy the angle ABC—specify carefully how it is to be copied—in order to obtain a line which, by Proposition 21 (I.27), we can conclude is parallel.)* □

Exercise 2

Identify where Proposition I.31 (see the previous exercise) is used in the proof of Proposition 24 (I.32).

Exercise 3

Make the proof of Proposition 24 (I.32) more formal by naming two points on the line through A parallel to BC, with A lying between these two points, so that all relevant angles can be clearly named.

Exercise 4

State and prove Proposition 21 (I.27) from memory.

Exercise 5

State and prove Proposition 22 (I.29) from memory.

Exercise 6

Distinguish between Propositions I.27 and I.29. What is assumed, and what is concluded? If you know the terms, use "necessary condition" and "sufficient condition."

Exercise 7

In the proof of Proposition 23 (I.30) we presumed the following statement. A line intersecting one of two parallel lines intersects the other one as well. We now prove this statement. Complete the proof.

Proof. Let AB and CD be parallel lines. Consider the line BE which intersects AB. Produce the line perpendicular to CD passing through B. Let it intersect CD in the point F... *(Show that angle ABF is right, by Proposition 22 (I.29). The lines BE and CD have a transversal, namely BG, intersecting them in such a way that, on one side, the interior angles are less than two right angles. Conclude by using the fifth postulate.)* □

Exercise 8

Produce a diagram to accompany the theorem of Thales, Proposition 25 (III.31).

6
The Composition of Quadrilaterals

6.1 Parallelograms

We have done three things so far. First, we established a clear foundation from which we build up all our explanations. This foundation includes both tangible experience (Chapters 1 and 2) and intelligible principles (Chapter 3). Second, we developed knowledge of the properties of triangles as they follow from this foundation (Chapter 4). Third, we examined parallel lines (Chapter 5).

We now proceed to discuss figures with four sides. We do not consider arbitrary four-sided figures. Instead, we restrict our attention to those in which opposite sides are parallel. We will see that they are composed of pairs of equal triangles. Here is the definition of the term we use.

Definition 26. *A parallelogram is a figure of four sides in which a pair of opposite sides is parallel and equal.*

The definition just given only asserts that there is a pair of opposite sides. We can in fact prove that the other pair of sides is also equal and parallel.

Recall something similar with isosceles triangles: only the sides are equal by definition, but we conclude that the angles are necessarily equal as well.

Proposition 27 (I.33)**.** *In a parallelogram, each pair of opposite sides is equal and parallel.*

Proof. Let $ABCD$ be a parallelogram, so that lines AB and CD are parallel and equal. Since AB and CD are parallel, we conclude by Proposition 22 (I.29) that angles ABC and BCD are equal.

Consider the triangles ABC and BCD. The angles ABC and BCD are known to be equal, the segment BC is common, and the segments AB and CD are equal by hypothesis. By Proposition 12 (I.4) we see that the two triangles are equal. In particular, the sides AC and BD are equal. The angles ACB and CBD are also equal, so that by Proposition 21 (I.27) the lines AC and BD are parallel. □

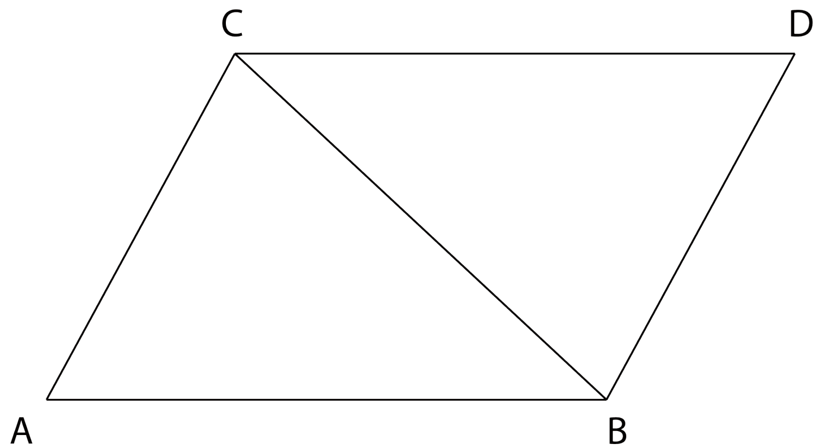


Figure 6.1: Propositions I.33 and I.34

The preceding proposition shows that parallelograms are uniform and symmetrical in a certain way. The next proposition shows how to relate such uniform shapes to things considered previously, triangles.

Proposition 28 (I.34)**.** *In a parallelogram, a diagonal divides the figure into equal triangles, and opposite angles are equal.*

Proof. Let $ABCD$ be a parallelogram, and let BC be a diagonal. The triangles ABC and BCD have three sides equal, by Proposition 27 (I.33) and the fact that BC is common. By Proposition 17 (I.8), then, the triangles are equal. In particular, angles BAC and BDC are equal.

The same reasoning applies to the diagonal AD. □

6.2 *Equality of Parallelograms*

The next proposition requires careful thought. It is something that might seem surprising. This could be the first result thus far where our conclusion is unexpected. One way to think about the surprising character of the result is this. This proposition says that if we know only that we are to enclose one acre of land with fence, we do not know how much fence we need. The amount needed could be arbitrarily large, depending on the shape of the piece of property.

Proposition 29 (I.35)**.** *Parallelograms on the same base between the same lines are equal.*

Proof. Let AB be the base, and let CD and EF be segments in a line parallel to AB such that AB, CD, and EF are all equal. Then $ABCD$ and $ABEF$ are each a parallelogram.

The triangles ACE and BDF are equal. This is seen as follows. Lines AC and BD are parallel and equal, by Proposition 27 (I.33). The angles at C and B are equal by Proposition 28 (I.34). The lines CE

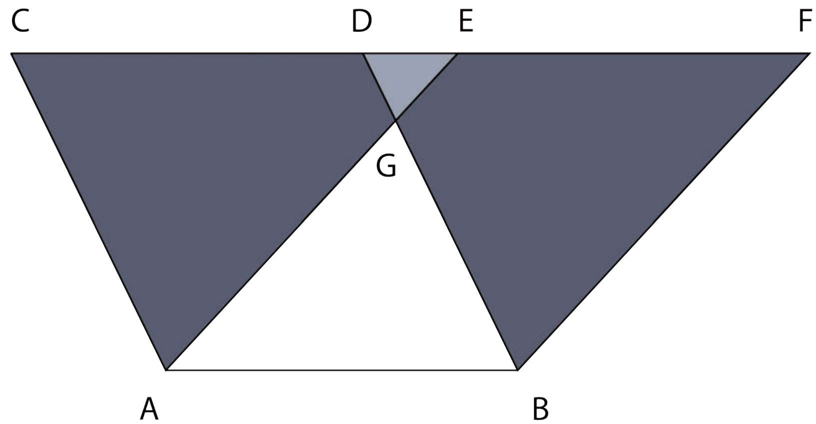


Figure 6.2: Proposition I.35

and DF are equal, since they are obtained by adding DE to the equal segments CD and EF. By Proposition 12 (I.4), then, triangles ACE and BDF are equal.

Let lines AE and BD intersect at G. The combination of the parallelogram $ABCD$ and the triangle DGE is the same as the combination of the triangle ACE and the triangle ABG. Similarly, the combination of the parallelogram $ABEF$ and the triangle DGE is the same as the combination of the triangle ABG and the triangle BDF.

Since triangles ACE and BDF are equal, we see that the combination of the parallelogram $ABCD$ and the triangle DGE is the same as the combination of the parallelogram $ABEF$ and the triangle DGE. Removing the common element DGE from both combinations, we see that parallelograms $ABCD$ and $ABEF$ are the same. □

It is important to observe that the word "equal" has been used in a somewhat different sense in this proposition. When we speak of equal triangles, it is something that we can imagine by direct movement or transposition. One triangle is placed, in our minds, upon the other, and they coincide. When we speak of equal parallelograms, we do not limit ourselves to those that coincide in this way via motion. We also include in the notion of "equal" those which decompose into triangles that are equal (in the stricter sense of equal used with triangles).

It is possible that you are familiar with the notion of "area" as a number associated to a shape. Using that term, we say that two parallelograms are equal when they have the same area. It is important to see, though, that our notion of equal does not involve any numbers. Instead, it involves breaking up a complex object into simpler objects.

We saw in Proposition 28 (I.34) that a triangle is half a corresponding parallelogram. This, in conjunction with the proposition we just proved on equality of parallelograms, allows us to conclude something about triangles on the same base. Keep in mind that the sense of equality used for triangles in the next proposition is the more general one that refers to decomposition into parts, rather than rigid translation, since it relies on Proposition 29 (I.35).

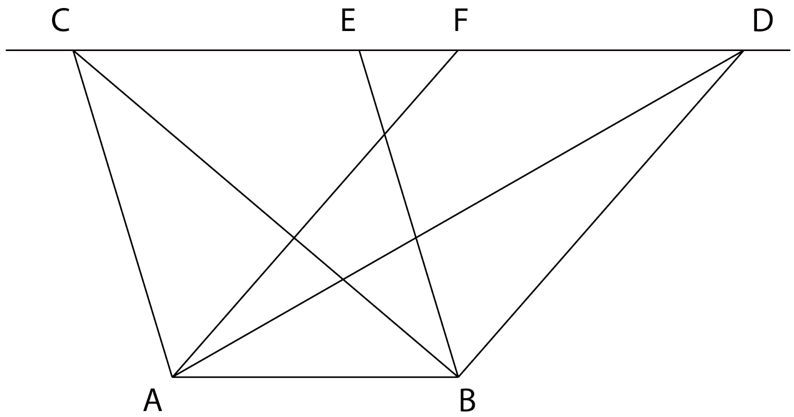


Figure 6.3: Proposition I.37

Proposition 30 (I.37). *Triangles on the same base and to the same parallel are equal.*

Proof. Let AB be the base, and let points C and D be taken on a line parallel to line AB. Let points E and F be taken on the line CD so that CE is equal to AB, and so that DF is also equal to AB. Then $ABCE$ and $ABDF$ are parallelograms. By Proposition 29 (I.35) we know that they are equal. Since triangle ABC is half of the parallelogram $ABCE$, by Proposition 28 (I.34), and triangle ABD is similarly half of the parallelogram $ABDF$, and the two parallelograms are equal, we see that the triangles ABC and ABD are the same. □

We have not used different words for the strict sense of "equal" (said only of triangles, and corresponding to rigid motion) and the looser sense of "equal" (said of both triangles and parallelograms, and corresponding to decomposition into strictly-equal parts). Keep this in mind when interpreting future statements. Take time to know clearly which sense of equality is relevant.

The next proposition is a kind of mixture of the two preceding.

Proposition 31 (I.41). *If a parallelogram has the same base as a triangle and is in the same parallel, then the parallelogram is twice the triangle.*

Proof. Let AB be the base, let CD be a segment equal to AB and parallel to AB, and let E be a point in the line CD. The quadrilateral $ABCD$ is twice the triangle ABC, by Proposition 28 (I.34). The triangle ABE is equal to the triangle ABC by Proposition 30 (I.37). Thus, the parallelogram $ABCD$ is twice triangle ABC. □

6.3 *The Pythagorean Theorem*

In order to understand our proof of the Pythagorean theorem, you must have a firm grasp of Proposition 29 (I.35) and its consequences.

Take the time to review that proposition carefully. Observe, further, how Proposition 31 (I.41) follows from it, by way of Proposition 30 (I.37).

The following terminology allows us to state the Pythagorean theorem clearly.

Definition 32. *In a triangle with a right angle, the side opposite the right angle is called the hypotenuse, and the other two sides are called legs.*

You showed in an exercise, using Proposition 20 (I.16), that a triangle cannot have more than one right angle. Thus, the word "hypotenuse" is unambiguous.

Definition 33. *A square is a four-sided figure with right angles and equal sides.*

If you study Euclid in greater detail in the future, you will see how he develops squares from propositions we have already given. At this time we will simply accept them, confident that they and their familiar properties are consequences of our postulates.

Theorem 34 (I.47)**.** *In a right triangle, the square on the hypotenuse is equal to the combination of the squares on the legs.*

We call this a "theorem" rather than a "proposition" to indicate its significance. There is no logical difference between a theorem and a proposition. Each is a mathematical statement with a proof.

Proof. Let ABC be a right triangle, and let angle ABC be right. Consider squares on each side of the triangle, whose vertices are as labelled in the figure. From the point B produce a line parallel to AD, intersecting AC in J in and DE in K.

Consider the triangles FAC and BAD. The sides FA and BA are equal, and the sides AC and AD are equal, being sides of squares. The angle FAC is equal to the angle BAD, since each is the combination of angle BAC with a right angle, and by Postulate 4 all right angles are equal. By Proposition 12 (I.4), the triangles FAC and BAD are equal.

The triangle FAC and the parallelogram $FAGB$ lie in the same parallels and on the same base, so $FAGB$ is twice FAC by Proposition 31 (I.41). The triangle BAD and the parallelogram $ADJK$ lie in the same parallels and on the same base, so $ADJK$ is twice BAD. Since the parallelograms $FAGB$ and $ADJK$ are twice things that are equal, they are themselves equal.

Reasoning in the same way with triangles IAC and BEC we can show that parallelograms $BCHI$ and $JKCE$ are equal.

To conclude, observe that the square on AC is the combination of the parallelograms $ADJK$ and $JKCE$.

□

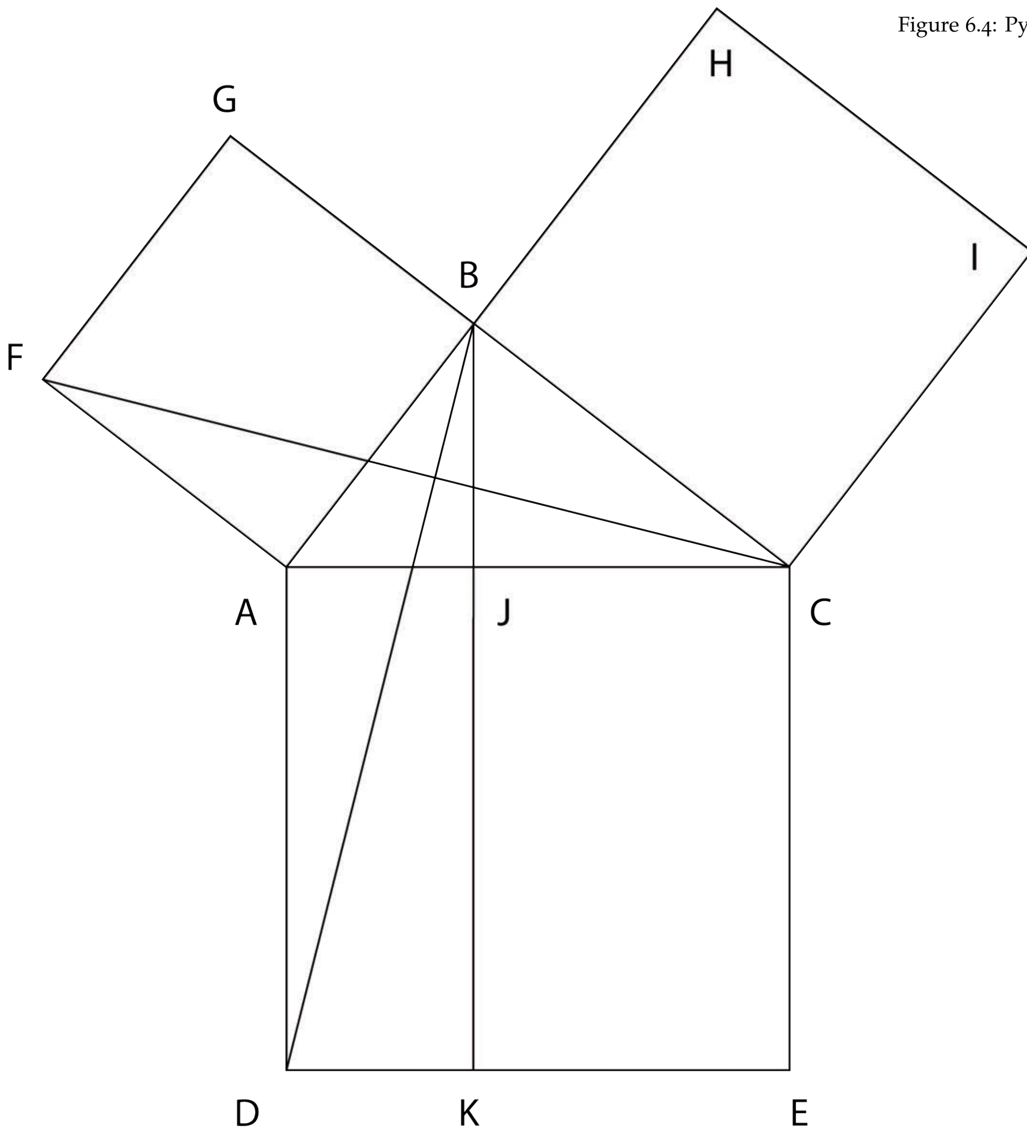


Figure 6.4: Pythagorean theorem

6.4 Exercises

Exercise 1

Draw a picture of a four-sided figure in which a pair of opposite sides is equal but not parallel. Is this a parallelogram? Explain.

Exercise 2

Draw a picture of a four-sided figure in which a pair of opposite sides is parallel but not equal. Is this a parallelogram? Explain.

Exercise 3

Prove Proposition 28 (I.34), that the diagonal divides the parallelogram evenly, using Proposition 12 (I.4) instead of Proposition 17 (I.8).

Exercise 4

Draw pictures of the three possible cases for Proposition 29 (I.35). One case is the one referred to in the proof and given in the figure, in which the two segments in the line share no common segment. Another case is when the two segments overlap. The third case is when they coincide exactly at an endpoint.

Exercise 5

Produce a suitable diagram, properly labeled, for Proposition 31 (I.41). Think about the possible locations of the point called E relative to the points called C and D.

Exercise 6

State and prove Proposition 29 (I.35) from memory.

Exercise 7

Our proof of the Pythagorean theorem mentioned Postulate 4, that all right angles are equal, for the first time since we listed the postulates. That postulate is also used implicitly in our proof. When constructing a square on one of the legs of the right triangle, we were confident that the side of the square produced on a leg was in the same line as the leg of the right triangle. One way to justify this is through Proposition I.14, which we have not yet seen. Complete the proof of that proposition below, giving a properly labeled figure as well.

Proposition 35 (I.14). *Let three rays emanate from a point, and suppose that the two adjacent angles together equal two right angles. Then the two outer rays are in a single line.*

Proof. Let AB, AC, and AD be the rays emanating from A, and suppose that the combination of BAC and CAD is equal to two right angles.

(We will reject this:) Suppose that B, A, and D are not in a line, and that instead there is a ray AE so that B, A, and E are in a line, and AE is distinct from AD. Then by Proposition 18 (I.13) angles BAC and CAE are two right angles. Since any two right angles are the same (here we use Postulate 4), angles BAC and CAE are together the same as angles BAC and CAD ...

(Remove angle BAC from each combination, and conclude that a part is the same as a whole, a contradiction.)

Therefore, B, A, and D are in a line. □

Exercise 8

Copy the proof of Theorem 34 (I.47).

Exercise 9

State and prove Theorem 34 (I.47) from memory.

Exercise 10

Proposition II.12 gives a statement like the Pythagorean theorem, but for obtuse triangles. To prove it, we will use a proposition of Book II, namely Proposition II.4.

Proposition 36 (II.4). *Let a segment be cut at random. The square on each piece, together with twice the rectangle on the two pieces, is the same as the square on the whole.*

Proof. A full proof requires more justification, but for now consider Figure 6.5, in which the whole square is shown to be the same as two smaller squares combined with two equal rectangles.

□

Using that proposition we can prove the following. Complete the proof, and give a properly labeled diagram.

Proposition 37 (II.12). *In an obtuse-angled triangle, the square on the side opposite the obtuse angle exceeds the squares on the remaining sides in the following determinate manner. Pick one of the acute angles, and from that vertex produce a perpendicular to the remaining side. This yields two segments in a line, namely the side of the triangle and the extension of the side to the perpendicular. The rectangle on these two segments is the requisite difference.*

Proof. Let ABC be the triangle with an obtuse angle at B, and let AD be the perpendicular through A to the line BC, with D in the line BC

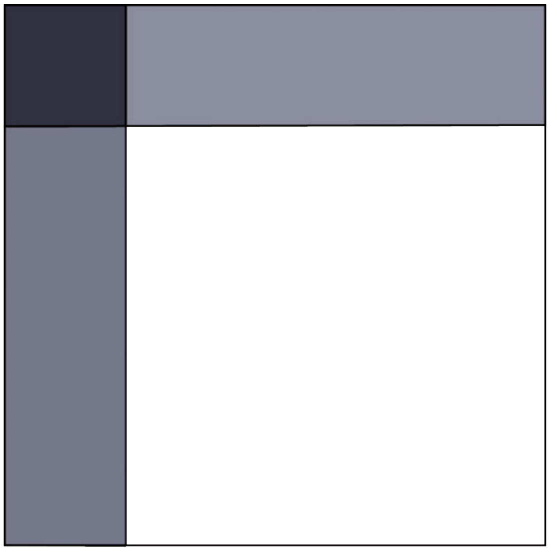


Figure 6.5: Proposition II.4

...*(Consider the right triangle ADC. Use Theorem 34 (I.47) to conclude something about squares associated to this triangle. Finally, use II.4, the proposition given above, to relate the square on CD to other quantities.)* □

Exercise 11

Euclid's Proposition II.13 offers an account of the relationship between the square on the side opposite an acute angle in a triangle, and the squares on the other two sides. Attempt to formulate a reasonable statement yourself. Do not simply look up Euclid's Proposition II.13. Compare to the previous exercise.

7
Ratio

7.1 Repetition and Comparison

Until this point, we have not made use of numbers, except in counting the points of a triangle and parallelogram. We will now use numbers to talk about the relationships that geometric objects have among themselves.

Let us begin with segments. If we are given two segments, we can compare them to each other to see which is larger. We can also place copies of a segment in a row, all in the same line and adjacent to one another. These two notions—comparing things and copying things—are at the heart of the idea of something called "ratio."

Definition 38. *A ratio is a relationship of size between two things of the same kind.*

This definition is hard to understand. We come to understand more of what it means by saying when two ratios are the same. The first step to this is to compare two ratios. We use symbols in the definition because there are many things involved.

Definition 39. *Let A and B be two things having a ratio, and let X and Y be another two things having a ratio. We say that the first ratio is less than the second ratio if there are natural numbers m and n so that m copies of A is less than n copies of B while m copies of X is greater than or equal to n copies of Y.*

Natural numbers are 1, 2, 3, 4, ...

Note that A and B need to be the same kind of thing, and X and Y need to be the same kind of thing, but that the two pairs can involve different kinds of things. You will see examples of this in the exercises.

Now we can give the definition of sameness for ratios. The first form is negative. That means it talks about what does not hold.

Definition 40. *Two ratios are the same if neither ratio is less than the other.*

To say that one ratio is less than another means that specific natural numbers m and n are found which satisfy the stated condition. To say that the ratio is not less than the other, then, means that no such natural numbers can be found. Therefore, the definition, reformulated without reference to comparison of ratios, makes an assertion about all pairs of natural numbers. It says that no pair that distinguishes the two ratios can be found. Here is a reformulation of the definition along those lines.

Definition 41. *(Equivalent to previous) Let A and B be two things having a ratio. Let X and Y be two other things having a ratio. We say that the ratio of A to B is the same as the ratio of X to Y if for every pair of natural numbers m and n, m copies of A compare to n copies of B in the same way that m copies of X compare to n copies of Y.*

By saying "compare in the same way" we mean that when m copies of A are bigger than n copies of B, m copies of X are bigger than n copies of Y, and that when m copies of A are smaller than n copies of B, m copies of X are smaller than n copies of Y. Finally, that if m copies of A are the same as n copies of B, then m copies of X are the same as n copies of Y.

Notation: Given things A and B which have a ratio, we will sometimes use the symbolic expression $A : B$ to refer to the ratio of A to B.

The idea of ratio presented here is due to someone named Eudoxus, who lived almost a century before Euclid. The concept is very powerful, but it is also challengingly abstract. Be patient, and work through the exercises on ratios of specific geometrical objects. They will help you to grasp ratio.

7.2 Ratio and Proportion in Plane Figures

We now have a definition of ratio that we understand reasonably well through some concrete exercises. A definition alone, though, is not especially useful. We need to develop tools by which we can conclude that ratios of a certain kind exist in specific cases. We need what we could call a "theory" or a "science" of ratio. That is what we develop now.

It will be difficult to read beyond this point if you have not yet done some of the exercises.

One goal in this section will be to show that, given two triangles satisfying a certain relationship, the ratio of a side of one triangle and a side of another triangle is the same ratio as the ratio of a different side of the first triangle and a different side of the second triangle. This is Euclid's Proposition VI.4. We do not prove this directly, by considering ratios of sides alone. Instead, we proceed by way of intermediate objects of a different kind than segments, namely, triangles.

The first proposition relating sides in triangles and the triangles themselves relies heavily on Proposition 29 (I.35) and the immediate consequence for triangles in Proposition 30 (I.37). Return to those statements and review them before continuing.

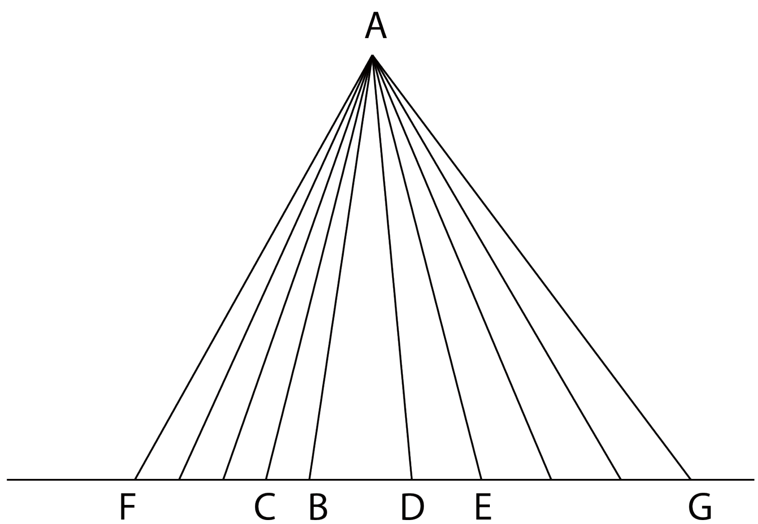


Figure 7.1: Proposition VI.1

Proposition 42 (VI.1). *Triangles between the same parallels have the same ratio as their bases. Parallelograms between the same parallels have the same ratio as their bases.*

Proof. Let triangles ABC and ADE lie under A with BC and DE in the same line. Let BC be copied m times, and let DE be copied n times, to obtain larger triangles ABF and ADG as in Figure 7.1. Each of the small triangles in the large triangle is equal to ABC or ADE, respectively, by Proposition 30 (I.37). The triangle ABF will exceed the triangle ADG if and only if m copies of ABC exceed n copies of ADE. The latter condition holds if and only if m copies of BC exceed n copies of DE.

This reasoning extends to the situation in which a second line passes through A, parallel to line BC, and the vertices of the triangles are not both the same (as they were here both A), but instead each is simply taken to be somewhere on this parallel line.

The statement for parallelograms follows from the statement that a parallelogram is twice the triangle obtained from a diagonal, Proposition 28 (I.34). One thing exceeds another exactly when half of the first exceeds half of the second. □

We will be able, using the preceding proposition, to show that triangles have a certain ratio provided that segments have that same ratio. The next proposition allows us to conclude that two segments have to one another the same ratio as two other segments.

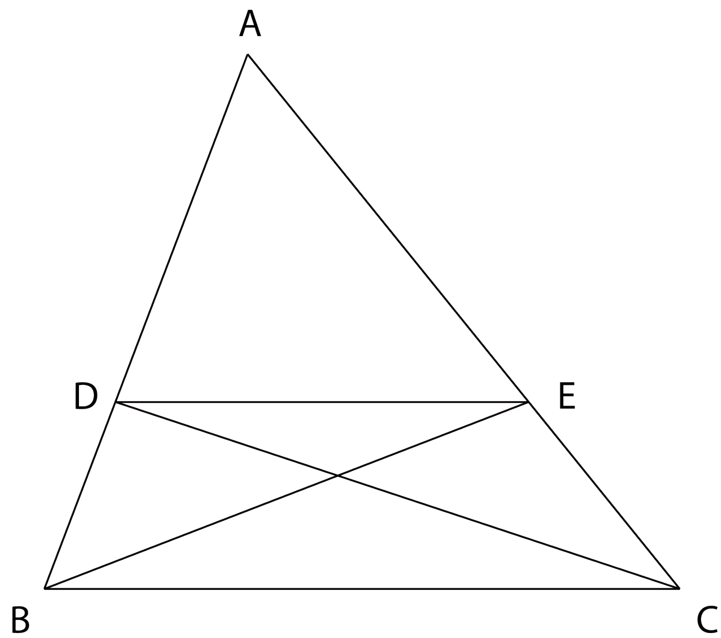


Figure 7.2: Proposition VI.2

Proposition 43 (VI.2)**.** *Let a line parallel to the base of a triangle intersect the two sides. The ratio of one piece to the other piece on one side is the same as the ratio of one piece to the other on the other side.*

This phrasing is terse. The precise ratios are made clear in the figure and the proof.

Proof. Let ABC be a triangle, and let the line DE be parallel to BC as in Figure 7.2. The triangles BDE and DEC are equal, in the more general sense of equality of decomposition. This is because they are on the base DE, and have vertices in the parallel line BC, so that Proposition 30 (I.37) applies.

Since ADE and BDE are triangles in the same line (AD) under the same point (E), the ratio they have with each other is the same as that of their respective sides AD and BD, by Proposition 42 (VI.1). Similarly, the ratio of the triangles ADE and DEC is the same as the ratio of their sides AE and CE. Since the triangles BDE and DEC are equal, and ADE is common to both ratios, we conclude that the ratio of AD to BD is the same as the ratio of AE to CE. This is what was to be shown. □

7.3 *Similarity and Right Triangles*

The preliminary, more technical propositions we have proved now enable us to demonstrate the following feature of equiangular (also called "similar") triangles.

Proposition 44 (VI.4)**.** *Let two triangles have equal angles, and let the sides of the triangles be paired with respect to the correspondence of angles.*

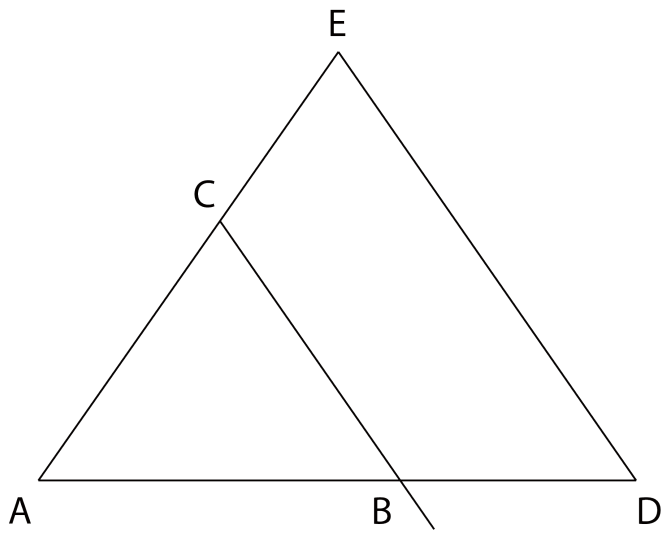


Figure 7.3: Proposition VI.4

Then the ratio of the side of one triangle to the side of the other triangle is the same in each case.

Proof. Let two such equiangular triangles be given, and produce on the same side of a segment of the larger triangle a copy of the smaller triangle, so that they share the angle. Proposition 16 (I.7) ensures the uniqueness of this copy. This yields nested triangles ABC and ADE as in Figure 7.3.

The angles ABC and ADE were presumed equal. Extending the line CB past B, and considering the vertical angle to ABC, we can infer that the lines BC and DE are parallel, using alternate interior angles and Proposition 22 (I.29).

Equality of vertical angles is Proposition 19 (I.15)

According to Proposition 43 (VI.2), the ratio of AB to BD is the same as the ratio of AC to CE. We now wish to show that, additionally, the ratio of AB to AD is the same as the ratio of AC to AE.

This is a separate proposition in Euclid. It is Euclid's V.17.

(We will reject this.) Suppose that the ratio of AB to AD is not the same as the ratio of AC to AE, and without loss of generality let the first ratio be the lesser.

Then there are natural numbers m and n so that m copies of AB are less than or equal to n copies of AD, while m copies of AC are greater than n copies of AE. Observe that n copies of AD are n copies of AB combined with n copies of BD. By the condition given (removing the copies of AB), $m - n$ copies of AB are less than or equal to n copies of BD. Reasoning similarly and considering AE as combined from AC and CE, we see that $m - n$ copies of AC are greater than n copies of CE. Such a pair of natural numbers $m - n$ and n, though, would prove that the ratio of AB to BD is less than the ratio of AC to CE, which contradicts the hypothesis that these

Note that since AE is larger than AC, m is greater than n.

two ratios are the same.

Therefore, the ratio of AB to AD is not less than the ratio of AC to AE, and by reasoning similarly we see that it is not greater. The same argument applies to the pairs of sides at each of the other angles similar to each other. □

Right triangles have a special property. The perpendicular from the right angle to the hypotenuse divides the triangle into two smaller right triangles. It turns out that these smaller right triangles are similar to the larger whole right triangle.

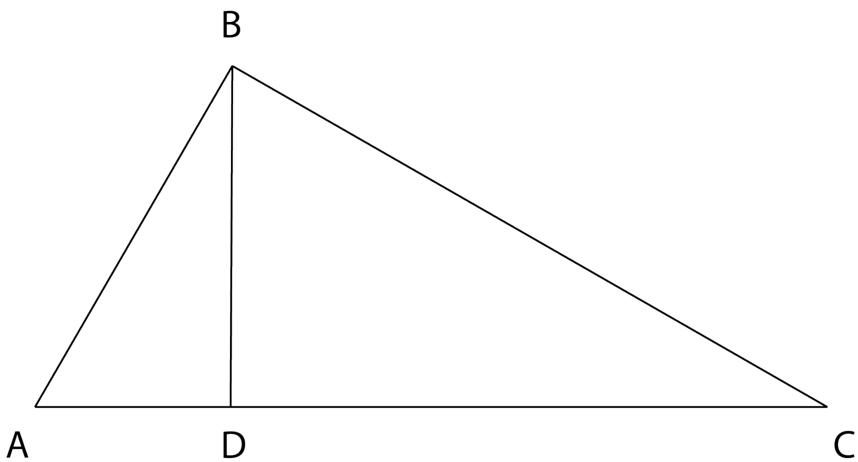


Figure 7.4: Proposition VI.8

Proposition 45 (VI.8). *If a perpendicular is dropped in a right triangle from the right angle to the hypotenuse, the resulting smaller triangles are similar to the whole, and to each other.*

The perpendicular from the right angle to the hypotenuse is the same line that we use to prove Theorem 34 (I.47). Go back to review this.

Proof. Let ABC be a right triangle with the angle ABC right, and let line BD be perpendicular to AC so that the triangles ADB and CDB are right. The angles in a triangle are two right angles by Proposition 24 (I.32). This means that angles DAB and ABD combine to make one right angle (consider the right triangle ABD) and that angles DAB and BCD combine to make one right angle (consider the given right triangle ABC). Since the combination of DAB with ABD is one right angle, and this is the same as the combination of DAB with BCD (which is also one right angle), we remove the common angle DAB to conclude that angles ABD and BCD are the same. This shows that the triangles ABC and ADB are equiangular, or similar, by Proposition 44 (VI.4).

Arguing in the same way, the triangles ABC and CDB are equiangular, and so similar. Finally, since two triangles equiangular to a third are also equiangular to each other, we see that the two smaller triangles ADB and CDB are similar. □

The following proposition is only a portion of a statement in Euclid, but it is enough for our purposes. The statement might appear somewhat unusual to you, so it is good to relate it to things that you already know. In more modern terminology it says that when the dimensions of a figure are rescaled by a certain factor, the area is rescaled by the square of that factor. Consider, for example, the area of a field whose length and width are three times the length and width of another field. The area of the larger field is nine times the area of the smaller.

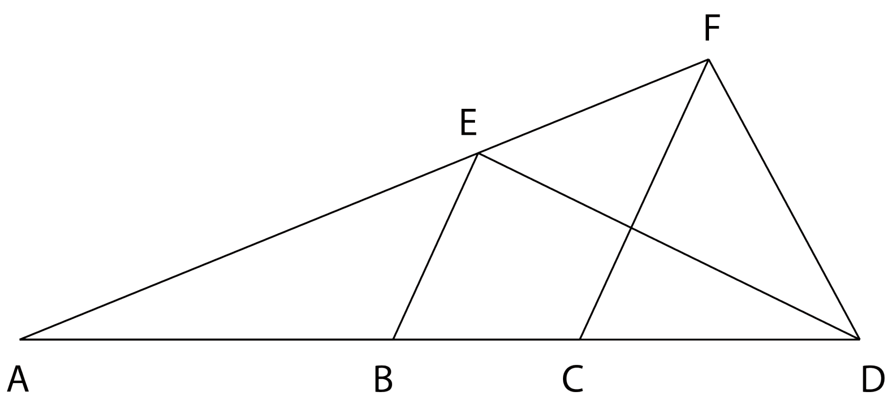


Figure 7.5: Proposition VI.19

Proposition 46 (VI.19 Porism)**.** *Let three lines be given, so that the ratio of the first to the second is the same as the ratio of the second to the third. Let similar triangles be produced on the first and second line. The ratio of the triangle on the first line to the triangle on the second line is the same as the ratio of the first line to the third line.*

Proof. Let the points A, B, C, and D be in a line, and let the lines AC and AB have the same ratio as lines AD and AC. A point E not on the line yields triangles ABE and ADE. Let a line through C be drawn parallel to BE, intersecting line AE at F. We then have triangles AFC and AFD.

1. $AC : AB$ is $AD : AC$	hypothesis
2. $AF : AE$ is $AC : AB$	VI.4 applied to triangles ACF and ABE
3. $FAD : EAD$ is $AF : AE$	VI.1 applied to point D and triangles on line AF

4. $FAD : EAD$ is $AC : AB$	Claims 2 and 3 (two things the same as a third are the same as each other)
5. $FAD : FAC$ is $AD : AC$	VI.1 applied to point F and triangles on line AD
6. $FAD : FAC$ is $AC : AB$	Claims 1 and 5
7. EAD is the same as FAC	By Claims 4 and 6, EAD and FAC have the same ratio to FAD. Things having the same ratio to one thing are the same.
8. $EAD : EAB$ is $AD : AB$	VI.1 applied to point E and triangles on line AD
9. $FAC : EAB$ is $AD : AB$	Claims 7 and 8, and the fact that things that are the same have the same ratio to another thing

□

Step 7 is Euclid's V.9. It is natural enough that we will just accept it. You might enjoy trying to prove it. Use proof by contradiction, assuming one thing greater than the other. Consider using Euclid's V.8, given in the exercises.

The previous proposition treats the ratio between the triangle on the middle segment and the triangle on the shortest segment. The same reasoning also applies, though, to the ratio between the triangle on the longest segment and the triangle on the middle segment.

The following generalization of the Pythagorean theorem was discovered by Euclid himself. This is what we are told by Proclus, who wrote a commentary on Euclid's *Elements*. Our figure only shows similar triangles on each side of the right triangle, but the theorem applies more generally to figures of many sides, such as quadrilaterals and pentagons and hexagons.

Theorem 47 (VI.31)**.** *The combination of similar figures on the legs of a right triangle is equal to the similar figure on the hypotenuse.*

Proof. Let a right triangle ABC be given with angle ABC right, and let D be produced on AC so that BD is perpendicular to AC.

The triangles ABD, CDB, and ABC are all similar, by Proposition 45 (VI.8).

From this we find that the ratio of AC to AB is the same as the ratio of AB to AD. By Proposition 46 (VI.19) we see that the ratio of the figure on AC to the figure on AB is the same as the ratio of AC to AD.

The proof is extended from triangles to arbitrary polygons by decomposing the polygons into triangular pieces. Thus, it suffices throughout this proof to read "triangle" for "polygon."

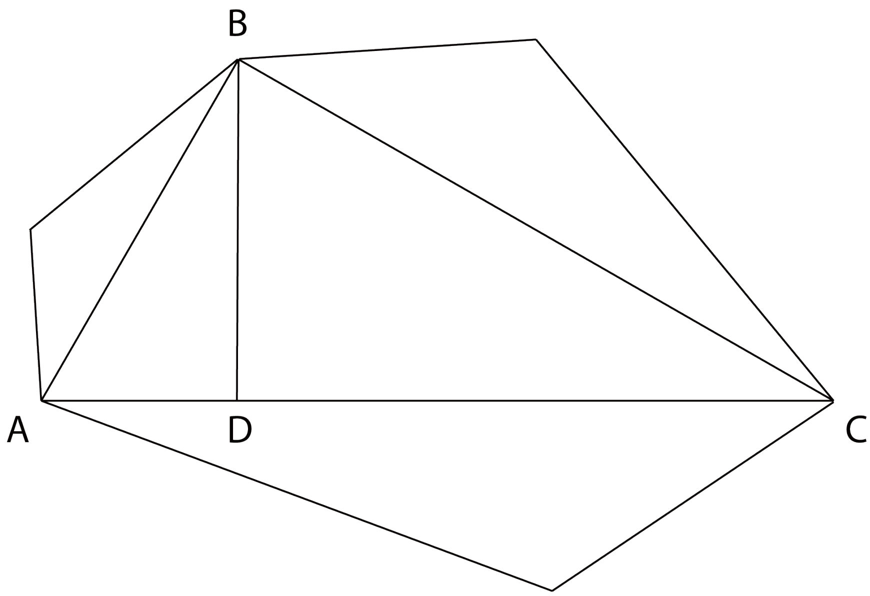


Figure 7.6: Theorem VI.31

Reasoning in the same way, we find that the ratio of the figure on AC to the figure on BC is the same as the ratio of AC to DC.

In this situation, since the parts AD and DC together make up the whole AC, it is possible to combine ratios to the parts (the manner of such combination is spelled out in the exercises). The conclusion from such a combination is that the ratio of the figure on AC to the combination of figures on AB and BC is the same as the ratio of AC to the combination of AD and DC. Since the combination of AD and DC is the same as AC, the second named ratio is the ratio of identity (of a thing to itself). Therefore, the equal ratio, of the figure on AC to the combination of figures on AB and BC, is also the ratio of identity.

This establishes the relationship among the figures on the sides of a right triangle that was to be shown.

□

7.4 *Exercises*

Exercise 1

Restate the definition of one ratio being less than another, using different letters. Let things F and G have a ratio, and let other things J and K have a ratio. What does it mean to say that the ratio of F to G is less than the ratio of J to K?

Exercise 2

Refer to Figure 7.7. Show that the ratio of segment A to segment B is less than the ratio of segment X to segment Y in the following way.

1. Produce a segment about four times as long (it doesn't need to be exact) as A. Copy off segment A three times, starting from one end. Mark the point which corresponds to three copies of A. Label it $3A$.

2. On the same long line, copy out segment B two times and mark the point. Label it $2B$.

3. Observe that three copies of A are less than two copies of B. (If you didn't get this, go back and start over.)

4. Make a different long line, this one about four times as long as X.

5. On this second line, copy out C three times, and mark the point. Label it $3X$.

6. On the second line, copy out Y two times, and label the point as $2Y$.

7. Observe that three copies of X are greater than two copies of Y.

8. Review the definition of ratio in Definition 39. Convince yourself that you have shown that the ratio of A to B is less than the ratio of X to Y.

Exercise 3

Refer to Figure 7.7. Show that the ratio of segment X to segment Y is less than the ratio of segment C to segment D in the following way.

1. Produce a segment about five times as long (it doesn't need to be exact) as X. Copy off segment X four times, starting from one end. Mark the point which corresponds to four copies of X. Label it $4X$.

2. On the same long line, copy out segment Y three times and mark the point. Label it $3Y$.

3. Observe that four copies of X are less than three copies of Y. (If you didn't get this, go back and start over.)

4. Make a different long line, this one about five times as long as C.

5. On this second line, copy out C four times, and mark the point. Label it $4C$.

6. On the second line, copy out D three times, and label the point as $3D$.

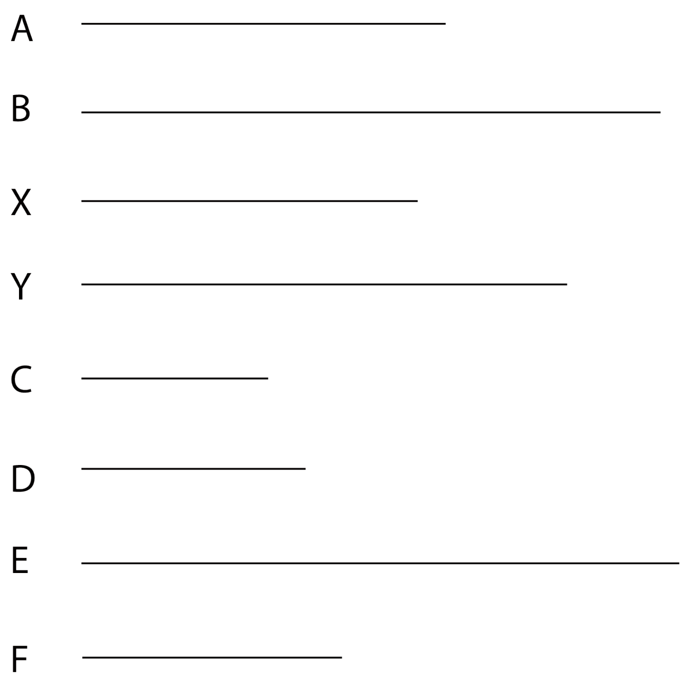


Figure 7.7: Segments for ratio exercises

7. Observe that four copies of C are greater than three copies of D.

8. Review the definition of ratio in Definition 39. Be careful about how the symbols are used. Their meaning in the definition is different than their meaning in this exercise. Convince yourself that you have shown that the ratio of X to Y is less than the ratio of C to D.

If you have difficulty with this step, use the equivalent definition that you created using different letters. Then see how those letters match up with X, Y, C, and D.

Exercise 4

Refer to Figure 7.7. Show that the ratio of Y to C is greater than the ratio of E to F.

- Note that showing that the ratio $Y : C$ is greater than the ratio $E : F$ is the same as showing that $E : F$ is less than $Y : C$.
- Use five copies of one (which one?) and two copies of the other to make the requisite comparisons.

Exercise 5

Refer to Figure 7.7. Show that the ratio of A to X is less than the ratio of F to D. You will need to use many copies of segments, but not more than ten of any segment.

Exercise 6

Carefully construct a square, and draw a diagonal in the square. Show that the ratio of the diagonal of your square to the side of your square is greater than the ratio of F to C (where those segments are in Figure 7.7). Use the numbers 7 and 5.

Exercise 7

Refer to Figure 7.8. Show that the ratio of the region A to the region B is less than the ratio of the segment C to the segment D. For convenience, regions equal to A within B have been marked. (Compare two copies of one thing with one copy of the other thing.)

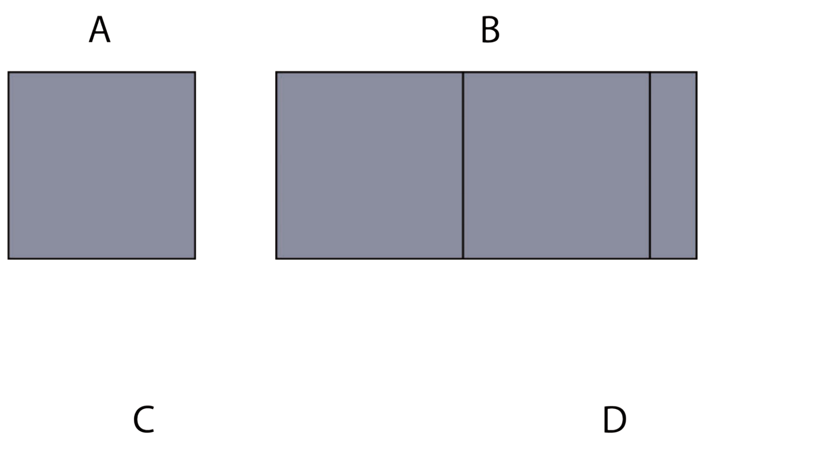


Figure 7.8: Rectangles and segments for ratio exercises

Exercise 8

Refer to Figure 7.9. Let A be the shaded region within the large square, and let B be one of the small squares (there are four of them) within the large square. Let C be the long line, and let D be the short line. Show that the ratio of A to B is less than the ratio of C to D. (Compare four copies of one thing to one copy of the other thing).

Exercise 9

Refer to Figure 7.10. Let A be the shaded region within the hexagon, and let B be the whole hexagon. Let C be the shaded region within the triangle, and let D be the whole triangle. Show that the ratio of A to B is greater than the ratio of C to D. (Consider six copies of one thing and one copy of the other thing.)

Exercise 10

Produce a square, and draw its diagonal. Show that the ratio of the diagonal to the side of the square is greater than the ratio 7 : 5. Note

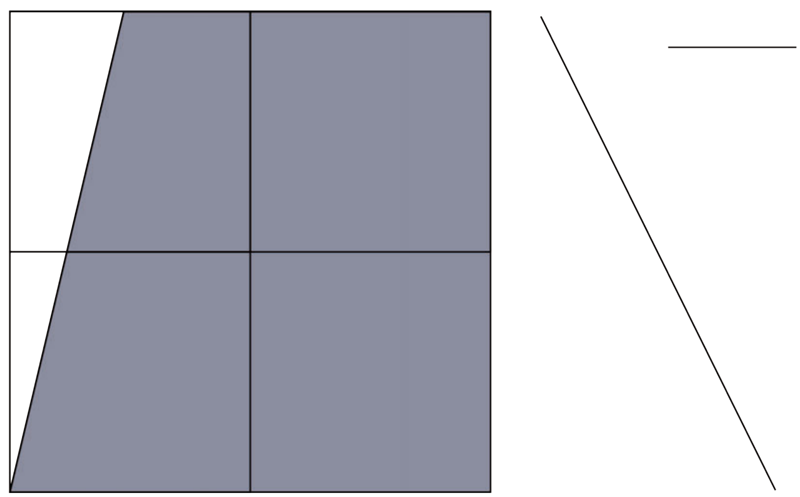


Figure 7.9: Trapezoid and segments for ratio exercises

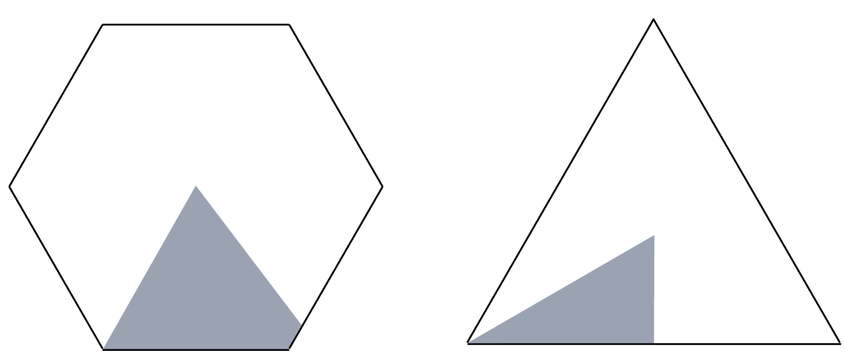


Figure 7.10: Regions for ratio exercises

that five copies of 7 is the same as seven copies of 5. Consider five copies of the diagonal in comparison with seven copies of the side of the square.

Exercise 11

Produce an equilateral triangle, and bisect one of the angles. Extend the angle bisector until it intersects the opposite side of the triangle. Show that the ratio of the side of the triangle to the bisector is greater than the ratio 8 : 7 but smaller than the ratio 7 : 6.

Exercise 12

Give a proof.

Proposition 48 (V.8). *Given three things, with the first greater than the second, then the ratio of the first to the third is greater than the ratio of the second to the third.*

Hint: Consider the difference of the greater and the lesser thing. Some multiple of it eventually exceeds the third thing. Use this to find a case when a given number of copies of the greater exceeds some number of copies of the third, but that number of copies of the second does not.)

Here we make an important assumption, namely that all the things we deal with are comparable via numbers. The formal mathematical term for this is *Archimedean*, a reference to the mathematician Archimedes.

Exercise 13

Proposition V.10 is the *converse* of Proposition V.8. To say that it is the converse means that the direction of implication is reversed. For an example of propositions related as converses, see Proposition 14 (I.5) and Proposition 15 (I.6), which deal with isosceles triangles. Using the fact that it is the converse of V.8 (given above), state Proposition V.10.

We also have another pair of converse propositions, involving parallel lines. See if you can find those two.

Exercise 14

Prove Proposition V.10. (See the previous exercise for the statement of the proposition.)

Exercise 15

Complete the proof below.

Proposition 49 (V.22). *Let six things of the same kind be given, and let them be separated into two groups of three. Suppose that the ratio of the first and second in the first group is the same as the ratio of the first and second of the second group. Suppose further that the ratio of the second and third in the first group is the same as the ratio of the second and third in the second group. Then the ratio of the first and the third in the first group is the same as the ratio of the first and the third in the second group.*

Proof. Let A, B, and C be the first group, and let D, E, and F be the second group. Then the ratio $A : B$ is the same as the ratio $D : E$, and the ratio $B : C$ is the same as the ratio $E : F$. Consider any numbers m, n, and p. Then m copies of A and n copies of B compare in the same way that m copies of D and n copies of E do. Similarly, n copies of B and p copies of C compare in the same way that n copies of E and p copies of F do.

Suppose that m copies of A exceed p copies of C. Then the ratio of m copies of A to n copies of B is greater than the ratio of p copies of C to n copies of B, by Proposition 48 (V.8). Then, by the presumed equality of ratios, we see that the ratio of m copies of D to n copies of E is greater than the ratio of p copies of F to n copies of E. Thus, m copies of D will exceed p copies of F.

An important point is assumed here. Suppose that A and B have the same ratio as D and E. Then, given natural numbers m and n, m copies of A and n copies of B have the same ratio as m copies of D and n copies of E. Try to prove this.

Similarly, if m copies of A are less than p copies of C, then m copies of D will also be less than p copies of F.

(Complete the statement. It is almost complete. You must understand what has been shown but need not generate much more argument.) □

Exercise 16

The goal of this exercise is to prove the following proposition about ratio.

Proposition 50 (V.18, proportion *componendo*)**.** *Suppose that two things are given, and each is divided into two parts. Suppose further that the ratio of the first part of the first thing to the second part of the first thing is the same as the ratio of the first part of the second thing to the second part of the second thing. Then the ratio of the whole first thing to one of its parts is the same as the ratio of the whole second thing to its corresponding part.*

In our proof of Proposition 44 (VI.4) we showed that equality of ratios of wholes and parts implies equality of ratios of parts with parts. We continue to use that result here. Reread the proof of VI.4 to see that again and to understand the precise statement. That intermediate fact is a separate proposition in Euclid, Proposition V.17.

a.) By rereading the proof of VI.4, determine the precise statement of V.17. Give it in a form like that of the current proposition. (It is like the current V.18 but has different hypotheses.)

b.) Consider the proof of Proposition V.18 below. Produce a suitably labeled diagram to accompany it.

Proof. Let A and B be the two parts of the first thing, called C, and let D and E be the two parts of the second thing, called F. By hypothesis the ratio of A to B is the same as the ratio of D to E. We wish to show that the ratio of C to A is the same as the ratio of F to D. Suppose that this were not the case. Then there would be some different thing, G, other than D, so that the ratio of C to A is the same as F to G. The thing G is either greater than or less than D but is not equal to it. Suppose that G is less than D. Let H be that by which D exceeds G, so that D is the combination of D and H. Applying V.17, we conclude that the ratio of A to B is the same as the ratio of G to the combination of E and H. On the other hand, the ratio of A to B is the same as the ratio of D to E by hypothesis. Thus, the ratio of D to E is the same as the ratio of G to the combination of E and H. This is impossible, since D is greater than G and E is less than E and H combined, so that the first ratio is certainly larger than the second one. The same reasoning applies in the case that G is greater than D. The ratio of C to A must be the same as the ratio of F to D. □

c.) In the preceding proof we considered the ratios of C to A and F to D but not the ratios of C to B and F to E. Determine whether the proof is sufficiently general to prove the proposition as stated.

Exercise 17

Complete the proof of Theorem 47 (VI.31) through the following steps. Expressions such as fig.AC are to be read as "the figure on segment AC."

a.) Let things W, X, Y, and Z be given. Suppose that the ratio $W : X$ is the same as the ratio $Y : Z$. Convince yourself that the ratio $X : W$ is the same as the ratio $Z : Y$. This follows from the definition of ratio.

b.) From what is given we have fig.AC : fig.AB is the same as $AC : AD$. Use part a to rewrite this. We also have that fig.AC : fig.BC is the same as $AC : DC$. Leave that one in that order.

Look back to the proof to see why this is. We used similar triangles and Proposition 46 (VI.19).

c.) Use Proposition 49 (V.22) to show that the ratio of fig.AB to fig.BC is the same as the ratio of AD to CD. (fig.AC and AC will serve as the middle or second terms, occurring in both ratios, that drop out, leaving you with the ratio of the first and the third.)

d.) Use Proposition 50 (V.18) to show that the ratio of the combination of fig.AB and fig.BC to fig.BC is the same as the ratio of AC to CD. (The "wholes" are, in the first case, the combination of two figures, and in the second case the whole segment AC.)

The ratio of a first thing to itself is the same as the ratio of a second thing to the second thing.

e.) Use Proposition 49 (V.22) to show that the ratio of the combination of fig.AB and fig.BC to fig.AC is the same as the ratio of AC to AC. (Hint: Use V.22 again. In the first case, the intermediate term will be fig.BC. In the second case, the intermediate term will be CD. You will need to change the order of terms in a ratio again, like in part a.)

Exercise 18

Task: Given two segments A and C, produce a third segment B so that the ratio of A to B is the same as the ratio of B to C. (Pay close attention to the order of the terms.)
Procedure: Produce a circle whose diameter is equal to the combination of A and C, and let the point at which the two segments meet be D. Produce the line through D that is perpendicular to the diameter passing through D. The portion of this line between the point D and the circumference of the circle is the desired segment B.

Practice this procedure with pairs of segments of varying lengths.

Exercise 19

Prove the proposition.

Proposition 51 (VI.13)**.** *The procedure of the previous exercise yields a segment satisfying the condition.*

Use the fact that a triangle in a circle having a diameter as one of its sides is necessarily right. Then use Proposition 45 (VI.8).

Exercise 20

Produce a square equal to an arbitrary triangle in the following way. (This is roughly Euclid's Proposition II.14.)

1. Pick one vertex of the triangle and produce the line through it and parallel to the side of the triangle opposite this vertex.
2. Produce the line through the chosen vertex and perpendicular to the opposite side (or, equivalently, the newly produced line).
3. Use these lines to obtain a rectangle equal to twice the original triangle.
4. Find the midpoint of one of the sides of the triangle to obtain a rectangle equal to the original rectangle.
5. Apply the procedure of VI.13 (see the previous two exercises) to the two sides of this rectangle to obtain a square equal to the original triangle.

Exercise 21

Given a triangle and a segment, produce a rectangle on that segment, so that the rectangle is equal to the given triangle. (This is roughly Euclid's Proposition I.44.)

1. Pick a vertex (call it A) of the triangle and call the other two points B and C. Produce the line through A perpendicular to BC. Let D be the point at which the line intersects BC.

2. Find the midpoint of BC, and call it E.

3. Let the given segment be called FG. Extend the segment FG from F, going away from G. Mark the point H on this extension so that FH is equal to BE.

4. At F, and perpendicular to FG, produce a line equal to AD. Call the end of this line J.

5. At J, produce a line parallel to FH (and so perpendicular to FJ) and extend it on the same side of FJ as H.

6. At H, produce a line perpendicular to FH and on the same side as J. Let the intersection of this line and the line made in the previous step be K.

7. Produce the line GJ, which intersects HK. Call the point of intersection L.

8. Produce on FG, the given segment, a rectangle whose other side is the same as KL.

Exercise 22

Let a point be given, and a line that does not pass through this point. Call the point A. In the line consider arbitrary segments BC and DE. Let m be a natural number. Let F be a point so that C is between B and F and BF is equal to m copies of BC. Similarly, let n be a natural number, and let G be a point so that E is between D and G and DG is equal to n copies of DE. Suppose that triangle ABF is greater than triangle ACG. What can you say about the ratio of BF and CG? (This exercise has you think carefully about the way that Proposition 42 (VI.1) is proved.)

8
The Golden Ratio

Having considered ratios in general, we will now consider a special ratio. This special ratio is often called the golden ratio. We will give two accounts of the golden ratio. The first involves equality of parallelograms. The second involves ratio. In Euclid's *Elements* the ratio first arises from equality of parallelograms, and we will proceed in the same order.

8.1 Square and Rectangle

Given a segment, we seek to divide it a certain point so that a square and a rectangle are equal. More precisely, we have the following.

Task: Given a segment AB, find a point C between A and B so that the square on AC is the same as the rectangle determined by BC and AB.

While the task says nothing about ratios, it turns out that it produces the golden ratio. The segments, squares, and rectangles are shown in Figure 8.1. There is a straightforward procedure for dividing a segment in this way. It involves producing a perpendicular and then bisecting a segment.

Procedure: At A, produce the line AD perpendicular to AB so that AD and AB are the same. Find the midpoint E of AD. With center E and radius EB, produce a circular arc that intersects the line AD on the side of AB opposite the point D. Let the intersection point be F. The segment AF is the desired segment. With center A and radius AF find the point G on AB so that AG and AF are the same.

It is necessary to prove that the procedure given above does indeed divide the segment in the manner sought.

Proposition 52 (II.11)**.** *The procedure given above yields a division of AB so that the square on AC is the same as the rectangle on BC and AB.*

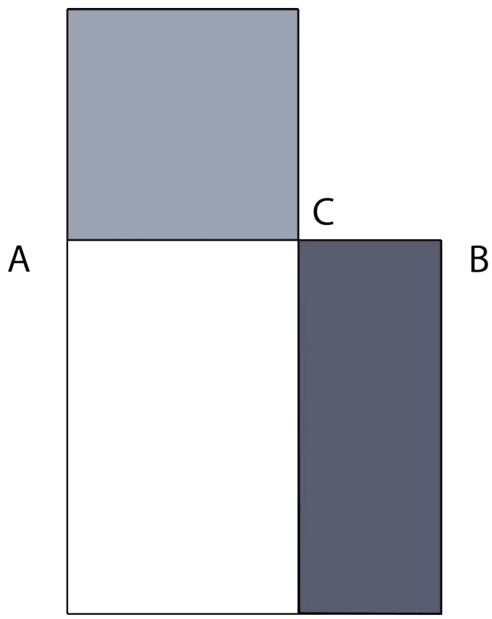


Figure 8.1: Square and Rectangle Equal

We have omitted much of Euclid's Book II, and so our proof must remain less formal and more diagrammatic. If you study Euclid in the future you will come to see this result in a broader light, and see how to make the proof more rigorous. Add labeled points to the given diagrams in order to check your comprehension of the proof.

Proof. Let the points be named as in the procedure stated above. The square on EB is, by the Pythagorean theorem, the same as the combination of the square on AE and the square on AB.

The square on AE together with the rectangle on AF and DF is the same as the square on EF. To justify this, see Figure 8.2, illustrating Euclid's Proposition II.6. The square on EF is broken up into four pieces, two of which are square and two of which are rectangular. The square on DE is the same as the square on AD, since AD and DE are equal. The little square (on DF) and the rectangle with side ED are both part of the long rectangle on AF. The two shaded regions correspond, since each is a rectangle with sides equal to AE and DF. Thus, we have shown the intermediate claim, that the square on AE together with the rectangle on AF and DF is the same as the square on EF.

The square on EF is the same as the square on EB, since EF and EB are radii of the same circle.

We then see, by means of the (equal) squares on EF and EB, that the square on AE together with the rectangle on AF and DF is the same as the square on AE and the square on AB. Removing common terms, the square on AB is the same as the rectangle on AF and DF.

Since AF and AC are the same, the rectangle on AF and DF is the same as the rectangle on AC and DF.

Removing the rectangle on AD and AC, the square on AC is the same as the rectangle on BC and AB. □

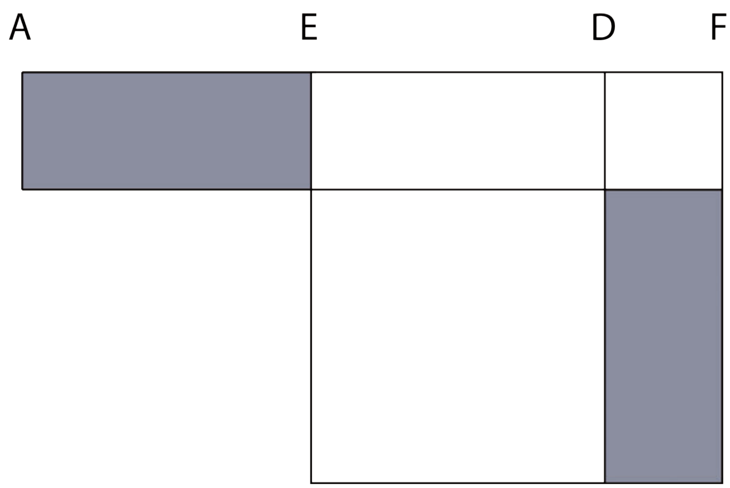


Figure 8.2: Euclid's II.6

8.2 *Another Perspective*

Let us now consider a different task. The task of the previous section was phrased in terms of sameness of parallelograms. This task is phrased in terms of ratio.

Task: Given a segment AB, produce a point C between A and B so that the ratio of AB to AC is the same as the ratio of AC to BC.

It turns out that the procedure of Section 8.1 also answers this question. In order to understand why this is so, we must make an observation about parallelograms which are the same and which have vertical angles.

Proposition 53 (VI.14, partial)**.** *In two parallelograms with vertical angles, if the shared lines are divided so that the ratio of one piece to the other of one of the lines is the same as the ratio of one piece to the other for the other line, in the opposite order, then the parallelograms are equal.*

Proof. See Figure 8.3. Let AB and CD intersect at E. The hypothesis is that the ratio of AE to BE is the same as that of CE to DE. Complete the sides to intersect in the point F, yielding the additional parallelogram EF.

By Proposition 42 (VI.1) the ratio of parallelograms AD and EF is the same as that of AE and EB. In the same way, the ratio of BC and EF is the same as that of CE and DE. Since the ratios of the divided

For convenience, we name the parallelograms by their diagonals.

segments are equal by hypothesis, we conclude that BD and AC have the same ratio to EF, which means that they are themselves the same. □

See the note regarding the proof of Proposition 46 (VI.19).

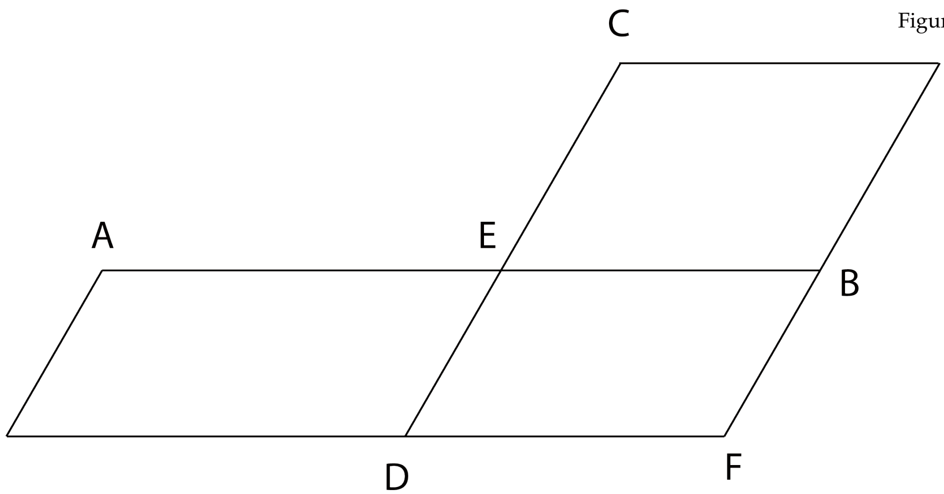


Figure 8.3: Proposition VI.14

The converse statement is also part of VI.14 and also holds. This means that if the parallelograms with vertical angles are equal, then the suitable ratios of sides are the same.

Now consider Figure 8.1 in the light of the preceding proposition. By the more elementary means of the previous section, we know that the two parallelograms in Figure 8.1, the square on AC and the rectangle on BC and AB, are equal. This means, by VI.14 (the second part, which we did not prove), that the ratio of AB to AC is the same as the ratio of AC to BC. The equality of parallelograms allows us to infer an equality of ratios.

8.3 A Remarkable Construction

You have already used the compass and straightedge to produce equilateral triangles and squares. Using the construction of the golden ratio, it is possible to produce a regular pentagon. We will present it only as a procedure. A proof that this procedure results in something regular requires a course going deeper into Euclidean geometry.

Task: Produce a regular pentagon.

Procedure: Draw a segment AB, and divide it at C in the golden ratio. Produce an isosceles triangle having two sides that are the same as AB and whose base is the same as AC. Circumscribe this triangle

You know a procedure for producing a circle passing through all three vertices of a given triangle.

with a circle. From the endpoints of the base of the isosceles triangle, make arcs with radius AC that intersect the circumscribing circle. Connect these two points to the triangle. The result is a regular pentagon.

As was noted above, to prove that this produces a regular pentagon requires further study of Euclid's *Elements*. Euclid's Book III considers circles and angles in them, and it is the key to understanding the procedure for producing a regular pentagon, which comes in Book IV. A rough outline of the proof is this. Having constructed the isosceles triangle whose sides and base have to each other the golden ratio, we can prove that the angles at the base are exactly twice the angle at the top. Then once we circumscribe the triangle, we bisect both of the larger base angles. The result is five different arcs of the circle, each corresponding to an equal angle situated at the circumference. All such arcs are equal, and the corresponding chords are as well.

8.4 Exercises

Exercise 1

Use a compass and straightedge to divide a given segment in the golden ratio.

Exercise 2

Produce a rectangle whose sides stand to each other in the golden ratio. Produce within this rectangle the square on the smaller side. This yields a square and a smaller rectangle. Within the smaller rectangle produce another square. Do this repeatedly.

Exercise 3

Produce an isosceles triangle whose two (longer, equal) sides have the golden ratio to the base. Produce an angle bisector of one of the base angles. This yields a small triangle with the base as one of its sides. Repeat the division, producing an angle bisector for one of the base angles of this (isosceles) triangle. Observe the similarity of all triangles produced.

Exercise 4

Produce a regular pentagon.

Exercise 5

Given a segment, find its midpoint. Construct on each half of the segment, on the same side, an equilateral triangle. Show that when the added points are connected, they produce an equal triangle. Use Proposition 24 (I.32).

Exercise 6

Produce a regular hexagon by repeating the construction of the previous exercise twice, once on each side of the line.

Exercise 7

Make a statement about the relationship between a regular hexagon and the circle circumscribing it (passing through all vertices). Refer to the previous exercise for an idea.

Exercise 8

Produce a regular fifteen sided polygon. To do this, produce an equilateral triangle and a regular pentagon both within a single circle, and sharing a vertex.

Exercise 9

An equilateral triangle is necessarily equiangular. What about quadrilateral (four sided) shapes with all sides equal? What about pentagonal (five sided) shapes?

Exercise 10

In the previous chapter you completed an exercise accomplishing this task.

Given a triangle and a segment, produce a rectangle on the segment that is equal to the given triangle.

Use Proposition 53 (VI.14) to justify the procedure. Use similar triangles to argue that certain ratios of segments are equal. Note that the equiangularity condition is not really important, since the procedure in question involves figures with right angles.

Exercise 11

If you have not yet done so, look ahead to the first chapter of Part IV: Astronomy. Make a schedule to complete the observational exercises. You should complete them while you are studying Parts II and III on Arithmetic and Music.

Part II

Arithmetic

Eloquia Domini, eloquia casta.

9 Counting

9.1 Words and Time

How long does it take to write a book? This is a complicated question. Books have various lengths. Some books are simple and others are complex. The question "How long?" can be answered with a number, but for this number to be sensible we must be clear about some of the decisions that we make along the way.

Books are made up of sentences, which themselves consist of words. One way, then, to get an idea of how long it takes to write a book is to see how many words a book has, and to see how long it takes to produce that many words.

Let us suppose we consider a book with 200 pages. How many words are in such a book? If you would like, you can take a book from your shelf and count how many words are on a page. Another thing you can do is this. Count how many words there are in one line, and count how many lines there are in one page.

I have done this with a book and gotten that there are about 9 words in a line and 28 lines on a page. That means that we need to consider the product 28×9 in order to find how many words are on a page. We find that there are roughly 252 words on a page. We now compute the product 252×200 to estimate the number of words that are in a book, since we chose the book to have 200 pages. This product is $50,400$. We can simplify things slightly by saying that such a book has about $50,000$ words, since 400 is small relative to $50,000$.

Now it is necessary to consider the rate at which words can be produced by an author. Let us consider typing. It is reasonable for a well-trained typist to produce at least 60 words per minute. Such speed, though, is attained only when copying text. An author must determine the words to be written, and this takes a significant amount of time. The typing is simply recording the words.

How do we deal with such a challenge? We could ask people who have written books or long articles. We could read biographies of writers. This might allow us to make a separate estimate for the time spent in thinking alone, and we could combine that with a different estimate made for the time used to record the thoughts in print. That is a reasonable approach, but in this case, for simplicity, we will simply combine the two kinds of authorial activity by choosing a slower rate of word production.

The rate of 60 words per minute corresponds to 1 word each second. It we consider an author spending time to think carefully about what is to be written, we can slow down the rate by a factor of 10. That means each word takes 10 seconds to produce, on average, and that about 6 words are produced each minute, or 360 each hour.

How many hours, then, will it take to produce 50,000 words? Here we can use division. The quotient of 50000 by 360 is 138 with a remainder. That means that it takes about 138 hours to write the book.

How much time can be devoted to writing each day? For some authors this might be 8 hours, for others only one, for others 2. Let us suppose the author has 3 hours each day for writing. The quotient of 138 by 3 is 46. Thus, we conclude that an author might reasonably write a book in 46 work days, or about two months.

Some prolific writers have written this rapidly.

Many books take much longer than this to write. Recall that we adjusted the rate of word production downward by a factor of 10 to account for thinking about what to say. What would happen if we adjust downward by a factor of 100? This extra factor of 10 will follow throughout the computation, yielding an estimate of about 460 work days to write the book. That works out to about a year and a half, allowing for days off from time to time.

This is a better general estimate.

9.2 *Principles for Estimation*

The first section gave an example in which each choice for the estimate was stated clearly. In the exercises that follow, you are to make similar choices. It is not important to make the guesses in an exact fashion. Often there is no such thing as an exact answer. Instead, there is a reasonable range of possibilities, and you are to get a rough idea of this range.

Here are some questions to ask yourself when counting things.

- How many of the things are in a small piece of the whole?
- How many pieces make up the whole?
- What is something familiar to me that is related to this?
- If I were to adjust this by a factor of 2 (or of 10), would it be better to adjust it upwards or downwards?

Do not bog down. The operations of arithmetic are exact, and you will use them, but there are many aspects of these questions that are not exact. You must move lightly and freely through them, taking enough time for reasonable choices but not lingering too long.

9.3 *Exercises*

Exercise 1

How many hairs are on your head?

Exercise 2

How many blades of grass are in your yard or the nearest park?

Exercise 3

The Library of Alexandria had about 100,000 books. Could one person hope to read them all?

Exercise 4

Could your home function as the Library of Alexandria (i.e., is there enough room)?

Exercise 5

How much water does your family use in one year?

Exercise 6

How long would it take you to dig, by hand, a hole as large as your room? As your house?

Exercise 7

Suppose that for one year all of your meals were cooked using only wood. How many trees would you use?

Exercise 8

How many stars can you see at night? (This depends on where you live. It also depends on how much time has passed since sunset, and the time of the year.)

The next exercises relate to our first example.

Exercise 9

Pick a book from a shelf. Pick what seems to be a representative line (not too short, not too many big words or too many short words). Count the words in that line. Count the lines on that page. Use those two to estimate the words on the page. Now count the words on page. Do this with a page from three different books. Are any of your estimates exact? Are they overestimates? Underestimates?

Exercise 10

Pick a book that you like, and choose a page within that book. Set a timer for one minute and type as much of the page as you can. Count how many words your were able to type. Repeat the exercise, but writing the words (neatly) by hand. These tests give you an idea of how many words you can produce per minute if the ideas are already prepared.

Exercise 11

Have someone pick one of the topics below for you. Immediately write for two minutes in response to the prompt. At the end of two minutes see how many words you have produced.

1. Describe the difference between day and night.
2. Describe the four seasons.
3. Describe boiling water.
4. Describe your earliest memory.
5. Describe your favorite book.
6. Describe your favorite food.
7. Describe a flower.

Exercise 12

Consider the topics in the previous exercise that you did not write about. How many words could you write about those other topics in two minutes?

10
Numbers in Themselves

In the first chapter we worked with numbers freely, seeing how they are used to count various objects. Now we will begin to consider properties that they have in themselves, without reference to material objects. We work with what are called "natural numbers," which means those that arise through counting. The natural numbers are $1, 2, 3, \ldots$ and do not include things like negative numbers and fractions (also called "rational numbers").

10.1 Parity

The word "parity" refers to evenness and oddness. We will not use it often. It will come up towards the end of our study of arithmetic.

Often a group of people wish to form two equal teams to play a game. How can a large group rapidly form into two teams of the same size? Have each person find a partner. From each pair send one partner to one side and one to the other side. If everyone had a partner, the teams will be of the same size. If one person failed to have a partner, there is an uneven number of people and some additional arrangement needs to be made.

We now formalize the reasoning suggested by the procedure for making teams. There is an even number of people in the whole group exactly when two teams of the same size can be made by pairing up everyone.

Definition 54. *To say that a natural number is even means that there is a second natural number such that the first number is twice the second one.*

It is important that we restrict our attention to natural numbers, rather than working with rational numbers.

Consider some examples. Is the number 6 even? Yes, because 3 is a natural number, and 6 is the same as 2×3. Is the number 16 even? Yes, because 8 is a natural number, and 16 is the same as 2×8.

Note that the definition of even does not have the form *When we divide it by 2 …* Statements like that involve our activity, so they are not appropriate for giving a definition of a mathematical thing. The number 6 is even whether or not we compute anything.

We can also define odd numbers.

Definition 55. *To say that a natural number is odd means that there is another natural number such that the first number is one less than twice the second.*

Every natural number is even or odd. It is possible to prove this using a certain principle, but we will take it as something given.

10.1.1 Power of 2

Among the even numbers there are certain special ones. These are the numbers obtained through multiplying repeatedly by 2. The first few such numbers are $2, 4, 8, 16, 32, 64, 128, 256, 512, 1024$. You will explore properties of this sequence of numbers in the exercises.

10.2 Triangular

Sometimes a collection of objects can be arranged in a triangular figure. There are three examples in Figure 10.1. Such an arrangement has nothing to do with the nature of the objects themselves. It is something that depends only on the number. When a number corresponds to such an arrangement it is called a triangular number. We see from the figure that 3, 6, and 10 are triangular numbers.

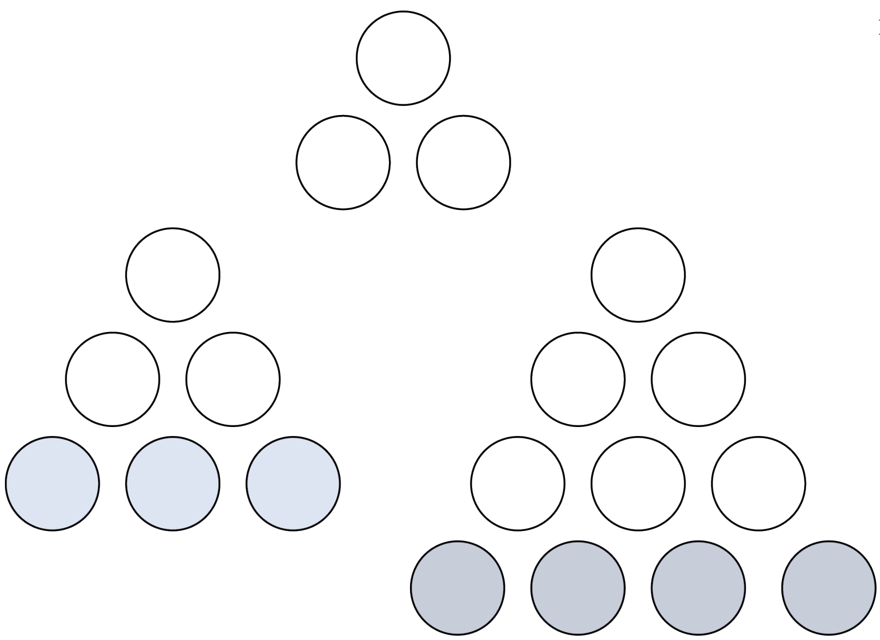


Figure 10.1: Triangular numbers

What is a way to produce a triangular number from another triangular number? Take the given triangular number and consider that many objects arranged in a triangle. Add one more row of objects.

This row has one more than the last row of the previous configuration. This is depicted in the figure. The small triangle has three objects. The next one is obtained by adding another row of three objects. The largest is obtained by adding yet another row, containing four objects.

10.3 Prime

Twelve objects can be arranged into a variety of rectangular configurations. Two are shown in Figure 10.2. Eleven objects, on the other hand, cannot be arranged into multiple equal rows. This corresponds to the fact that no number less than eleven divides it.

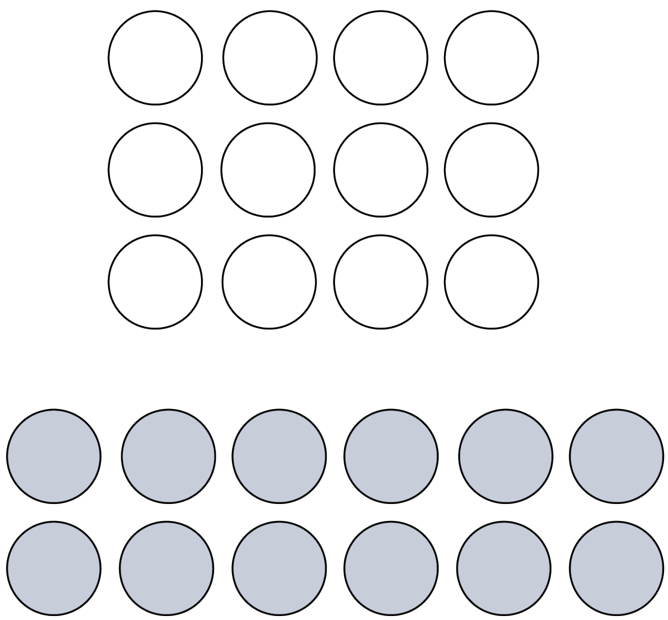


Figure 10.2: Rectangular configurations

When a number, like eleven, does not allow for rectangular arrangements, we say it is prime. More formally, instead of referring to rectangles, we speak of divisibility. We will define the term "divisibility" carefully later, but for now you can think of it in the sense that is familiar to you.

Definition 56. *A number larger than* 1 *is said to be prime if it is divisible only by 1 and itself.*

There is a good reason to exclude the number 1 in the definition of prime. For now, do not worry about it too much. The numbers 2, 3, and 5 are the first three prime numbers.

10.4 Perfect

The number 10 is divisible by 1, 2, 5, and 10. The sum of the numbers 1, 2, and 5 is 8, which is less than 10.

The number 12 is divisible by 1, 2, 3, 4, 6, and 12. The sum of the numbers 1, 2, 3, 4, and 6 is 16, which is greater than 12.

The number 6 has a special property. It is divisible by 1, 2, 3, and 6, and the sum of the numbers 1, 2, and 3 is 6. We say that 6 is a perfect number.

Definition 57. *To say that a natural number is perfect means that it is the sum of its proper divisors.*

The word "proper" in "proper divisor" means that the number itself is excluded.

10.5 Exercises

Exercise 1

Memorize the first 10 powers of 2.

Exercise 2

Find the following sums.

a.) $1 + 2 + 4$

b.) $1 + 2 + 4 + 8$

c.) $1 + 2 + 4 + 8 + 16$

Guess what this sum is, based on your previous computations.

$$1 + 2 + 4 + 8 + 16 + 32 + 64 + 128 + 256$$

Check your guess, either by hand or by using a calculator.

Exercise 3

Consider this row of consecutive powers of 2.

$$2 \quad 4 \quad 8 \quad 16 \quad 32$$

There are five terms. The third one, 8, is in the middle. Observe that the square of the middle term is equal to the products of the terms on the end. In other words, 8^2 is the same as 2×32.

Pick another series of an odd number of consecutive powers of 2. See that the same relationship holds between the middle and the end terms.

Exercise 4

Consider this row of consecutive powers of 2.

$$2 \quad 4 \quad 8 \quad 16 \quad 32 \quad 64$$

This contains an even number of terms, so unlike the previous exercise, there is no term exactly in the middle. The two that are closet to the middle, though, are 8 and 16. Their product is 128, which is also the product of the end terms 2 and 64.

Pick another series of an even number of consecutive powers of 2. See that the same relationship holds between the central and end terms.

Exercise 5

If instead of "natural number" we say "integer" (which also includes $0, -1, -2, \ldots$) we could make a different definition of odd numbers, saying that a number is odd when it is the successor of an even number (rather than one less than it). Show that these two definitions are equivalent. That means, show that anything falling under the first falls under the second, and vice versa.

Exercise 6

List all prime numbers less than 30.

Exercise 7

Make pictures of all ways of arranging 18 objects into rectangles. Observe that there are a number of points of ambiguity that you should resolve. Does a width of one unit count as yielding a rectangle? What about if rectangles are the same under reflection or rotation? State clearly how you resolve these ambiguities in order to assert that you have enumerated all possible arrangements.

Exercise 8

For each number from 4 to 30 which is not prime (you know the primes from an earlier exercise), determine whether the number is perfect. For numbers that are not perfect, note whether they are larger or smaller than the sum of their proper divisors. (There is one more perfect number other than 6 in this range.)

No prime can be perfect, since the sum of the proper divisors of a prime is simply 1. Primes do not have any proper divisors other than 1.

11
Demonstration with Natural Numbers

11.1 Even Numbers

Review the definition of *even number* given in the previous chapter. It asserts that a number is even when there is another number that corresponds to the size of a team when the whole group is divided into two teams.

The next proposition records the familiar fact that the sum of two even numbers is also even. The purpose of this proposition is not to present to you something new, something that you did not know. Instead, the purpose is to accustom you to proofs using natural numbers. You must attend to, and use, the specific definition we have made about evenness.

Proposition 58. *The sum of two even numbers is even.*

Proof. Let m and n be the even numbers. Since m is even, there is a natural number j so that m is the same as $2j$. Since n is even, there is a natural number k so that n is the same is $2k$. Then the sum

$$m + n$$

is the same as

$$2j + 2k$$

and this number can also be expressed as

$$2(j + k)$$

which shows that the sum $m + n$ is indeed twice another natural number, namely the natural number $j + k$. □

What is our goal? We have chosen the names m and n for two even numbers. Our goal, if we wish to show that the sum is even, is to show that there is some number such that $m + n$ is the same as 2 times that number. You might write this out like this.
Goal: Show that $m + n$ is the same as 2(something).

We extract the common factor of 2.

We have arrived at the goal we set out to reach. The "something" in question is the number $j + k$.

The word "sum" is used for addition, and the word "difference" is used for subtraction. We now see that a similar result holds for differences of even numbers.

Proposition 59. *The difference of two even natural numbers is an even number.*

Proof. Let m be the greater even number and n be the lesser even number. Since m is even there is a natural number j so that m is $2j$. Since n is even there is a natural number k such that n is $2k$. Then the difference

Goal: Show that $m - n$ is the same as 2(something).

$$m - n$$

is the same as

$$2j - 2k$$

which is the same as

$$2(j - k)$$

which shows that $m - n$ is twice a natural number, namely the number $j - k$. □

We have arrived at the goal.

We can also investigate the relationship of even numbers and the operation of multiplication. In this case we will not assume that both numbers are even. It is enough if one of them is. The proof explains why.

Proposition 60. *The product of an even natural number with any natural number is even.*

Proof. Let m be the even natural number and n be another natural number. Since m is even, there is a natural number k so that m is the same as $2k$. Then the product

Goal: Show that mn is the same as 2(something)

$$mn$$

is the same as

$$(2k)n$$

which is the same as

$$2(kn)$$

The product of a group of numbers is independent of how we pair them when we multiply two at a time.

so that the product mn is even, since it is twice a natural number, namely the number kn. □

We have arrived at the goal.

11.2 Odd Numbers

Recall the definition made earlier of what an odd number is. A number is said to be odd when it is one less than twice some number, i.e., one less than an even number.

Proposition 61. *The sum of two odd numbers is even.*

Proof. Let m and n be the two odd numbers. Since m is odd, there is a number j so that m is the same as $2j - 1$. Similarly since n is odd there is a number k so that n is the same as $2k - 1$. Then

Goal: Show that $m + n$ is the same as 2(something).

$$m + n$$

is the same as

$$(2j - 1) + (2k - 1)$$

which is the same as

$$2j + 2k - 2$$

which is the same as

$$2(j + k - 1)$$

so that the sum $m + n$ is indeed even. □

We have arrived at the goal.

The statement about products of odd numbers has a slightly more complicated proof. Go through slowly, step by step.

Proposition 62. *The product of two odd numbers is odd.*

Proof. Let m and n be the two odd numbers. Then there are numbers j and k such that m is the same as $2j - 1$ and n is the same as $2k - 1$. Then mn is the same as

$$(2j - 1)(2k - 1)$$

which is the same as

$$4jk - 2j - 2k + 1$$

which is the same as

$$2(2jk - j - k) + 1$$

and this number is odd, since it is the same as

$$2(2jk - j - k + 1) - 1$$

. □

Observe that to conclude we must make clear that the definition is satisfied, and the definition we have established for odd numbers means being one less than, rather than one more than, an even number.

In the conclusion of the proof above, we had to be creative in order to exhibit that the number satisfied the definition of odd. Essentially we did this. There was an even number $2a$, and we wished to show that $2a + 1$ is odd. We considered the next even number $2a + 2$, which can also be written as $2(a + 1)$. Then the number $2a + 1$, the odd number after $2a$, is the same thing as $2(a + 1) - 1$, the odd number before $2a + 2$.

We do not need to give a proposition about the product of an even and an odd number. We already did so when we showed that the product of an even number with any number (odd or not) is even, in Proposition 60.

11.3 Divisibility

We now formalize the notion of divisibility. Our method generalizes what we did for evenness. A number is even if it is divisible by 2, according to the definition that follows.

Definition 63. *To say that a first number divides a second number means that the second is a multiple of the first number. We also say that the second number is divisible by the first number.*

We say that 3 divides 12, since 12 is the same as 3×4. Another way to phrase this is that 12 is divisible by 3. Observe that the definitions of even and divisibility together mean that evenness is the same as divisibility by 2.

You must keep in mind the distinction between the terms "divides" and "is divisible by".

Here is how the definition is used in proofs involving divisibility. We are given two numbers, one of which divides the other by hypothesis. We then give names to the numbers, for example we give the name d to the number that divides the other, and the name m to the number that is divisible by d. The definition then asserts that we conclude that a third number exists, namely the number of times d is repeated in order to yield m. We will give a name to that number, say k. Then we will be able to say that m is the same as dk.

Proposition 64. *The sum of two numbers divisible by a third number is also divisible by that third number.*

Proof. Let m and n be numbers that are both divisible by the number d. Then there is a number j so that m is dj and a number k so that n is dk. Then the number

Goal: Show that $m + n$ is the same as d(something).

$$m + n$$

is the same as

$$dj + dk$$

which is the same as

$$d(j + k)$$

so that the sum $m + n$ is divisible by d. □

We have reached the goal, since $m + n$ is now shown to be d times something. That something is $j + k$.

Look back to the proof of Proposition 58. The proof we just gave is almost exactly the same. It is slightly more general. Rather than considering the specific number 2, we worked with a general divisor d. The coming proposition is similarly a generalization of a statement we already saw about even numbers.

Proposition 65. *A number that divides a second number also divides every multiple of the second number.*

Proof. Let the number d divide the number n, and let m be any number. Since d divides n there is a number k so that n is the same as dk. Then mn is the same as

Goal: Show that mn is the same as d(something).

$$m(dk)$$

which is the same as

$$d(mk)$$

and thus mn is divisible by d. □

The size of the parts is given by considering the same multiple, multiplying by m.

The next proposition is a slight reformulation of the preceding one.

Proposition 66. *A number that divides a second number also divides all numbers divisible by that second number.*

Proof. Let d be a number that divides a second number m. Let n be a number that is divisible by m.

Goal: The number n is the same as d(something).

Since d divides m, there is a number j so that m is the same as dj. Since m divides n, there is a number k so that n is the same as mk. The number mk, which is the same as n, is also the same as

$$(dj)k$$

which is the same as

$$d(jk)$$

and therefore n is divisible by d. □

We have shown that n is a multiple of d and thus reached the goal.

Here is a Euclidean proposition that relates our consideration of odd and even numbers to the more general study of divisibility.

The symbol IX is the Roman numeral for the number nine. This proposition is in the ninth book of Euclid's *Elements*.

Proposition 67 (IX.30)**.** *An odd number that divides an even number also divides half the even number.*

Proof. Let d be the odd number, and let n be the even number. Since n is even, there is a number k so that n is the same as $2k$. Since d divides n, there is a number j so that n is the same as dj. Therefore, $2k$ is the same as dj. The product of two odd numbers is odd, and d is odd, thus j must be even, since otherwise $2k$ would be odd, which is impossible. Thus, j is even, and there is a number ℓ so that j is the same as 2ℓ. Then dj is the same as

We are doing a quick proof by contradiction within the bigger proof.

$$d(2\ell)$$

which is the same as

$$2(d\ell)$$

which is the same as $2k$. We then conclude that

$$d\ell$$

is the same as k, so that d does divide k, and k is the half of n. □

11.4 Identifying False Statements

The mathematical statements that we show to be true have the form "in every instance that certain conditions hold, we are justified in concluding some other thing." To show that a claim of this form is false, one only needs to show that there is a single instance in which the conditions hold yet the conclusion does not. Such an instance means that the universality asserted by the claim fails to hold.

Here is an example. We will clearly label the statement as false.

> **FALSE:** The product of two numbers, each of which divides a third number, also divides that third number.

This is false. The numbers 2 and 6 each divide the number 18, but the product of 2 and 6, namely 12, does not divide 18. The claim made is incorrect; we cannot draw the conclusion given there.

The numbers 2, 6, and 18 constitute what is called a "counterexample." A counterexample is an instance that shows that a claim that purportedly holds about all things of a certain form does not hold of one of them. The counterexample thereby exhibits that the claim is false.

Warning: It does not matter that there could be some instances in which the conditions as well as the conclusion are true. For example, the numbers 3 and 4 divide 24, and their product 12 also divides 24. This "example" does not constitute a proof of the false statement above. The important thing to keep in mind is that the statement asserts that in every instance that the conditions hold, the conclusion follows. It is one thing for the conclusion to follow from certain hypotheses, and another for it to hold accidentally.

A proof of the mathematical statement explains how the conclusion follows necessarily from the assumed conditions. It is not sufficient to give an example of an instance in which the conditions and the conclusion are both true.

11.5 Axioms for Natural Numbers

When we introduced Euclidean geometry in the first part of this book, we made certain postulates clear before proceeding. In this part, on arithmetic, we have begun to do demonstrations without making all principles fully explicit. The reasons for this are somewhat subtle. Let us consider, though, the sort of statements that could be axioms for natural numbers. We have relied on these and will continue to do so. Think about these statements and convince

In contemporary mathematical speech, the words "axiom" and "postulate" are interchangeable. Aristotle distinguishes between related Greek words, but it would be a mistake to make a direct comparison between his terms and our current ones. Aristotle's distinctions are reasonable, and modern mathematical practice is reasonable.

yourself that they are reasonable.

- Each natural number has a successor, a number that immediately follows it. Formulated symbolically, this asserts that given a natural number n, there is also a natural number $n+1$.
- The number 1 is not the successor of any natural number, but every other natural number is a successor.
- There is an addition operation for natural numbers.
- The addition operation is commutative, meaning that the sum is independent of the order of the summands. In symbols, given natural numbers m and n, the numbers $m+n$ and $n+m$ are the same.
- The addition operation is associative, meaning that given three terms, it does not matter which two we sum first. In symbols, given natural numbers k, m, and n, the numbers $(k+m)+n$ and $k+(m+n)$ are the same.
- There is a multiplication operation, and it is commutative and associative as well.
- The product of a number with a sum is the sum of the products of that number with each summand. In symbols, given natural numbers k, m, and n, the numbers $k(m+n)$ and $km+kn$ are the same. This relationship is referred to as distributivity.
- Every collection of natural numbers has a smallest element.
- The natural numbers are ordered by the successor relation (meaning we can speak of numbers "less than" and "greater than" other numbers), and the operations of addition and multiplication preserve this ordering (for example, $m > n$ implies $km > kn$ for any natural numbers k, m, and n).

Note that commutativity of multiplication is a Euclidean proposition (VII.16). Distributivity, in a certain sense, is Euclid's VII.5. These are things you might study if you consider Euclid's *Elements* in greater depth in the future. Whether a statement is an axiom or a proposition depends on what foundation has been assumed.

11.6 *Exercises*

Exercise 1

Use a diagram to show that $2(j+k)$ and $2j+2k$ are the same, for some chosen j and k. Use colors or symbols to clearly delineate teams of size j or k.

Exercise 2

The sum of an odd number and an even number is odd. Prove this. Note that you have not been given an exact proof of this. You have been given proofs of related statements. When you understand those proofs clearly you will be able to offer your own proof, using your own words.

Exercise 3

Prove that the difference of two odd numbers is even. Note that you have not been given an exact proof of this. You have been given proofs of related statements. When you understand those proofs clearly you will be able to offer your own proof, using your own words.

Another way to say "difference" is "the larger number minus the smaller number." It is important to interpret the statement using this order so that the difference of two natural numbers will itself be a natural number.

Exercise 4

The difference of two numbers that are divisible by a third number is also divisible by that third number. Prove this.

This generalizes a statement about even numbers whose proof is given in the text.

Exercise 5

Each statement is false. Find a counterexample by which to show that the statement is indeed false.

a.) If a first and second number divide a third number, then the sum of the first and second number also divides the third number.

b.) If one number divides a second number, then the square of the first number also divides the second number.

c.) If a first and second number divide a third number, then either the first divides the second, or the second divides the first.

d.) If a first number divides a second, and a third number divides a fourth, then the sum of the first and third numbers divides the sum of the second and fourth numbers.

Exercise 6

Determine which statements below are true and which are not. To show such a statement to be false, simply produce a single counterexample. To show it is true, give a proof.

a.) The square of an even number is divisible by 4.

b.) Let a first number and a second number each divide a third number. Then the first number divides the sum of the second and third numbers.

c.) Suppose that a first number divides a second one. Then the square of the first divides the cube of the second.

d.) The difference of two numbers, each of which divides a third, also divides that third number.

Exercise 7

Suppose a large group of people are together in a room, and many of them are from New Jersey. Someone says, "Everyone in this room is from New Jersey." Is this true? Is it false? Consider each of the following situations independently.

1. Person A says, "I'm from New Jersey."

2. Person B says, "I'm from New Jersey, and everyone here whom I've met is also from New Jersey."

3. Person C says, "I've never met someone who isn't from New Jersey."

4. Person D says, "I'm not from New Jersey."

Is everyone in the room from New Jersey? Which statement or statements allow you to determine whether the original claim (that everyone there is from New Jersey) is true?

Exercise 8

Explain the connection between the previous exercise and the counterexample given in Section 11.4, on *False Statements*. What is the thing that is like the word "everyone" in the statement about numbers? Or is the "everyone" only implicit?

Exercise 9

Prove that the product of three consecutive numbers is divisible by 6.

Exercise 10

Use the numbers 37, 15, and 4 to show that subtraction is not an associative operation. Write the numbers in that order, and consider the two ways in which you might choose to introduce subtraction and parentheses.

The subtraction operation is not defined (as a natural number) for all pairs of natural numbers, but it is defined whenever the first number is greater than the second.

Exercise 11

Use the numbers 100, 10, and 5 to show that division is not an associative operation. Just write 100, 10, and 5 in order, place division operators between each pair of sequential terms, and then consider the two ways in which you could introduce parentheses.

This is a tricky point. The division operation is not even **closed** with respect to natural numbers. There is not, for example, a natural number corresponding to $7 \div 3$.

Exercise 12

Translate the propositions of this chapter to "If... then..." statements, using symbols. For example, *If d divides m and d divides n, then ...*

12
Primes and Relative Primality

Recall the definition of prime number. A number larger than 1 is said to be prime when it is divisible only by 1 and itself, not by any other natural number. We are able to list small prime numbers fairy easily, using our familiarity with small numbers and their divisibility. The first six prime numbers are 2, 3, 5, 7, 11, and 13.

12.1 Sieve of Eratosthenes

Suppose that we wish to generate a list of all the prime numbers up to some fixed size. A mathematician named Eratosthenes discovered how to do this in a simple, systematic manner. His method is called the Sieve of Eratosthenes.

First, here is the general idea of the Sieve of Eratosthenes. A number that is not prime is a multiple of some smaller number. In order to get a list of primes, we get rid of all numbers that are multiples of smaller numbers. The numbers that remain must be prime.

Here is a verbal description. The Sieve of Eratosthenes proceeds in this way. Write a list of all the natural numbers up to some fixed bound. Cross off 1, because it isn't prime. The next number is 2. Circle it, since it has not been crossed off. It is prime. Now cross off all multiples of 2. After doing that, move on to the next number that isn't crossed off. (In this case, 3.) Circle it. It is prime. Cross off all its multiples. Move on to the next number that isn't crossed off. Circle it; it is prime. Cross off its multiples. Keep proceeding in this way until you reach a prime number bigger than the square root of the bounding value. (In the example using the table that follows, you can stop once you cross off multiples of 13, since the next prime is 17, whose square exceeds 200.) At this point, anything that hasn't been crossed off is a prime number.

You might find it clearer to follow the procedure as presented below.

1. Determine an upper bound, and list all numbers from 1 through that upper bound.
2. Cross off the number 1.
3. Circle the smallest number that is not crossed off.
4. Cross off all multiples of the largest circled number.
5. Find the smallest number that is not crossed off and not circled.
   (a) If the square of this number is less than the upper bound, go back to step 3 and continue.
   (b) If the square of this number is larger than the upper bound, stop. Circle all remaining numbers that have not been crossed off. The collection of circled numbers is all the primes between 1 and the upper bound.

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100
101	102	103	104	105	106	107	108	109	110
111	112	113	114	115	116	117	118	119	120
121	122	123	124	125	126	127	128	129	130
131	132	133	134	135	136	137	138	139	140
141	142	143	144	145	146	147	148	149	150
151	152	153	154	155	156	157	158	159	160
161	162	163	164	165	166	167	168	169	170
171	172	173	174	175	176	177	178	179	180
181	182	183	184	185	186	187	188	189	190
191	192	193	194	195	196	197	198	199	200

Figure 12.1: Table for Sieve of Eratosthenes

12.2 *The Number of Primes*

Loosely speaking, we can tell that it is "more difficult" for a larger number to be prime, since there are more possible factors. It is in fact the case that the prime numbers appear less frequently as numbers grow in size. It is nonetheless also the case that there are infinitely many prime numbers. This should not be thought to be obvious. It is a striking, significant fact. Moreover, it is justifiable in terms of the simple principles we have developed thus far.

12.2.1 *Preliminary Lemma*

Before attacking the problem of showing that prime numbers continue without end, we record a lemma, a helping statement, that we will use in the proof.

Lemma 68. *A natural number larger than 1 does not divide successive natural numbers.*

Proof. Let d be a natural number, and let n and $n+1$ be successive natural numbers. Suppose that d does divide both n and $n+1$. Then by the proposition proved in an earlier exercise, it would also divide their difference. The difference $(n+1)-n$ is simply 1. No natural number larger than 1 divides 1, so we obtain a contradiction. Thus, d divides at most one of n and $n+1$. □

How do we know that no natural number divides 1? It is a sufficiently clear statement that we can take it as given. Depending on what has been set down as the foundation for our arithmetic, we can also argue from the compatibility of multiplication with the ordering of the natural numbers. Multiples of a number are bigger than that number, and so multiples of a number bigger than 1 are also bigger than, and so not equal to, 1.

12.2.2 *Why can't we collect all primes?*

The proof that the primes continue indefinitely is somewhat demanding. Here is the structure of what we must do in order to complete the proof.

- Name an auxiliary quantity.
- Consider cases about which we reason separately.
- Draw a general conclusion using the conclusions in particular cases.
- Argue using proof by contradiction.
- Rise from a statement about a particular collection of prime numbers to a statement about all collections of prime numbers.

While studying the proof, identify when each listed item above is in play.

Theorem 69 (IX.20)**.** *The collection of all prime numbers is not finite.*

Proof. Consider a finite list of prime numbers $\{2, 3, 5, 7, \ldots\}$.

Multiply them all together and call the product M. Consider the number $M+1$. Either it is prime or it is not prime. *Case 1*: If $M+1$ is prime, it is not contained in the original list under consideration. The

Here we use the ellipsis to indicate that the list keeps going beyond the prime 7. We do not mean, however, that it continues without bound.

reason for this is that it is much bigger than all the numbers on that list, since M itself is a multiple of all of them. Thus, $M + 1$ is not on the list.

Case 2: Suppose that $M + 1$ is not prime. It must be divisible by some prime number; let's call that prime number p. The prime p could not have been on the original list, for the following reason. Suppose that it were. Then p would divide M. But p divides $M + 1$ as well, since that is how we defined p in the first place. According to Lemma 68 it cannot be the case that a prime number p divides both M and $M + 1$. If p were on the list, then, we would arrive at a contradiction. So p cannot be on the list.

In any case: Whether or not $M + 1$ is prime, we see that there is necessarily a prime number that is excluded from the original finite list.

The argument, moreover, applies to any finite list. We did not rely on any special properties of a specific list. Therefore, every finite list of primes omits some prime number, and so no finite collection includes all prime numbers. □

It is important to note that the two cases we considered are *comprehensive*. They include all possible situations.

12.3 Two Open Questions

All of the mathematics in this book has been known for centuries, even for millennia. This does not mean, though, that the study of mathematics is entirely closed or complete. Here are two places where mathematicians can still make new advances.

12.3.1 Twin Primes

Consider the following pairs of prime numbers.

- 3 and 5
- 5 and 7
- 11 and 13
- 17 and 19

What pattern do you observe? In each case, both numbers are prime, and they are also separated by 2.

The closest that two primes can be is 2 units, since in a pair of consecutive numbers one is necessarily even. The one exception is the pair of primes 2 and 3, since 2 is the sole even prime.

Definition 70. *Twin primes are primes that are separated by* 2.

We can then ask a question that is like the one answered by Theorem 69 (IX.20). Do the twin primes keep going?

Conjecture 71 (Twin Primes Conjecture)**.** *There are infinitely many pairs of twin primes.*

This is called a "conjecture." Recall that "conjecture" means that mathematicians think that the statement is true but have not been able to give a proof of it. At the time this book is being written, no one knows whether it is possible to prove the conjecture. Thus, no one knows how many twin primes there are.

12.3.2 Goldbach's Conjecture

With the exception of 2, all primes are odd. That means the sum of two primes different from 2 is necessarily even. We might try see whether all even numbers arise in this way. Here are some examples.

- 6 is the same as the sum $3 + 3$
- 8 is the same as the sum $3 + 5$
- 10 is the same as the sum $5 + 5$
- 12 is the same as the sum $7 + 5$
- 28 is the same as the sum $11 + 17$

After exploring many examples with small numbers, you might expect that each even number can be written as the sum of two primes.

Note that we should exclude 2, since we do not say that the number 1 is prime.

Conjecture 72 (Goldbach's Conjecture)**.** *Every even number larger than* 2 *is the sum of two primes.*

This conjecture occurs in correspondence in the 18th century between mathematicians named Goldbach (for whom the conjecture is named) and Euler.

Euler is a very important mathematician. If you have seen the number called e, related to something called the "natural logarithm," you have seen something named for him.

12.4 Greatest Common Divisors

When we discussed divisibility and primes, we considered numbers on their own. We now consider them in relationship with a partner number.

Definition 73. *The greatest common divisor of a pair numbers is the largest natural number that divides each number of the pair.*

Here are some examples. The greatest common divisor of 8 and 12 is 4. The greatest common divisor of 38 and 24 is 2. The greatest common divisor of 25 and 16 is 1. The greatest common divisor of 21 and 30 is 3.

It is possible to identify the greatest common divisor of one pair of numbers with the greatest common divisor of a related pair of numbers. This seemingly abstract statement will be quite useful computationally.

Proposition 74. *The greatest common divisor of a larger number and a smaller number is the same as the greatest common divisor of the difference between the larger and the smaller number, and the smaller number.*

Proof. Let m be the larger and n the smaller number, and suppose that d is their greatest common divisor. By an earlier exercise, we know that d divides $m - n$. Thus, d divides both $m - n$ and n, and so cannot be greater than the greatest common divisor of $m - n$ and n.

Go back and find the relevant propositions about divisibility. We rely on them here, even though there is no explicit reference.

Let k denote the greatest common divisor of $m - n$ and n. Since k divides $m - n$ and n, it also divides their sum, which is m. Since k divides m and n, it cannot be greater than d.

Since neither d nor k can be greater than the other, they must be the same. Thus, the greatest common divisor of m and n is the same as the greatest common divisor of $m - n$ and n. □

It might be helpful to consider a concrete example. The greatest common divisor of the numbers 70 and 28 is 14. The proposition asserts that 14 is also the greatest common divisor of 28 and 42, the latter of which is the difference $70 - 28$.

There is a special term for natural numbers whose greatest common divisor is 1. This means that the two numbers share no divisors.

Definition 75. *Two natural numbers are said to be relatively prime if their greatest common divisor is* 1*.*

It is important to observe that being *relatively prime* is a fundamentally different notion than being *prime*. While "prime" is said of a single natural number, "relatively prime" is said of a *pair* of natural numbers. Those numbers might themselves be prime or composite. For example, the number 8 is not prime, and the number 15 is not prime, but we say that 8 and 15 are relatively prime because they share no divisors, i.e., their greatest common divisor is 1.

12.5 Euclidean Algorithm

When the numbers being considered are small, we can find the greatest common divisor simply by looking. It is possible, though, to discover greatest common divisors in a methodical way, using the proposition of the preceeding section. This method is called the "Euclidean algorithm." An algorithm is a procedure for computing something.

Algorithm 76. *To find the greatest common divisor of a pair of numbers:*

1. *Find the largest multiple of the smaller number that is not greater than the larger number.*
2. *(a) If the smaller number divides the larger number (i.e the larger is a multiple of the smaller), stop. The smaller number is the greatest common divisor of the original pair of numbers.*

   (b) If the smaller number does not divide the larger number, go back to step 1. with the pair consisting of the smaller number and the difference between the larger number and the largest multiple of the smaller number.

Examples make the algorithm clearer. Here is an example of this algorithm using 48 and 33.

Initial pair: 48 and 33

Step 1: 33 is less than 48, but 2×33 is larger than 48, so 33 is the largest multiple.

Step 2b: 33 does not divide 48. We consider the difference $48 - 33$, which is 15. We go back to step 1 with the pair 33 and 15.

Step 1: 2×15 is less than 33, but 3×15 is greater than 33, so 2×15 is the relevant multiple.

Step 2b: 15 does not divide 33, so we consider the difference $33 - (2 \times 15)$, which is 3. We go back to step 1 with the pair 15 and 3.

Step 1: 5×3 is the largest multiple of 3 that does not exceed 15.

Step 2a: Since 15 is the same as 5×3, we stop. The last difference, 3, is the greatest common divisor of the original pair, 48 and 33.

In order to prove that the algorithm in fact produces what it is supposed to produce, it is helpful to have the following adaptation of Proposition 74.

Lemma 77. *The greatest common divisor of two numbers is the same as the greatest common divisor of the smaller number and the difference of the larger number and the largest multiple of the smaller number which is not greater than the larger number.*

Proof. Let m be the greater, and n the smaller, number. Let kn be the largest multiple of n which is not greater than m. By Proposition 74, the greatest common divisor of m and n is the same as the greatest common divisor of $m - n$ and n. If k is 1 we are done. Suppose k is greater than 1. Applying Proposition 74 to the pair $m - n$ and n, we see that the greatest common divisor of $m - n$ and n is the same as the greatest common divisor of $m - 2n$ and n. Continuing in this way, we see that the greatest common divisor of $m - kn$ and n is the same as the original quantity considered, namely the greatest common divisor of m and n. □

Proposition 78 (VII.1, VII.2). *The Euclidean algorithm yields the greatest common divisor of a pair of numbers.*

Proof. In the steps of Algorithm 76 above, at each point we go to step 1 with a pair obtained by subtracting the largest possible multiple of the smaller number from a larger number. By Lemma 77, the greatest common divisor of such a pair is the same as the greatest common divisor of the original pair. $\square$

12.6 Exercises

Exercise 1

Carry out the Sieve of Eratosthenes using the table provided on page 98 (Figure 12.1).

Exercise 2

Find a list of the first 1000 numbers, or make one, and use the Sieve of Eratosthenes to find all primes less than 1000. (Hint: once you circle 31 and cross off its multiples, you can stop.)

Exercise 3

In the Sieve of Eratosthenes, why can we stop once the square of the largest circled number exceeds the upper bound?

Exercise 4

The product $1 \times 2 \times 3 \times 4 \times 5$ is 120. The numbers $120 + 2$, $120 + 3$, $120 + 4$, and $120 + 5$ are all composite (i.e., not prime). We can check this directly, or we can proceed by an indirect route. Note that both 120 and 2 are divisible by 2, and so their sum is as well. Similarly, 120 and 3 are both divisible by 3, so that their sum is too. The same reasoning applies for 4 and 5.

Reflect on the remarks above and produce 10 consecutive numbers that are not prime. Can you find 100 consecutive numbers that are not prime? What does it mean to "find" such numbers? Can you write them down?

Exercise 5

The proof of Theorem 69 (IX.20), on the infinitude of primes, relied on the use of a number M, the product of all the primes in the collection under consideration. For each specific collection of primes below, compute M, and then determine whether $M + 1$ is prime or composite.

a.) $\{2, 3, 5\}$

b.) $\{2,7\}$

c.) $\{3,5,7\}$

d.) $\{2,3,5,7\}$

Exercise 6

Copy the proof that there are infinitely many primes.

Exercise 7

Give the proof of Theorem 69 (IX.20), that there are infinitely many primes, from memory.

Exercise 8

Find all pairs of twin primes between 100 and 200, using the table you produced with the Sieve of Eratosthenes.

Exercise 9

Pick three even numbers larger than 50. Write each one as a sum of two primes.

Exercise 10

Find even numbers that can be written as a sum of two primes in more than one way.

Exercise 11

Find the greatest common divisor of these small pairs using your familiarity with arithmetic.

a.) 10 and 25

b.) 21 and 35

c.) 60 and 48

d.) 45 and 39

Exercise 12

Find the greatest common divisor of each pair using the Euclidean algorithm.

a.) 98 and 126

b.) 156 and 96

c.) 1587 and 987

d.) 178 and 288

13
Linear Diophantine Equations

The greatest common divisor of two numbers can be used to understand the ways that those two numbers combine to form other numbers. This idea—a number being expressed as a combination of other numbers—is the theme of the present chapter.

13.1 Puzzles about Combination

13.1.1 Measuring Water

Consider the following puzzle. We are near a river, and we have a large empty bucket. We also have two glasses. One glass holds, when full, exactly 11 ounces. The other glass holds, when full, exactly 17 ounces. Neither glass has any markings on it. We only know the contents of each glass if it is exactly full. Using just these instruments—the river, the bucket, and the two glasses—can we put exactly 3 ounces of water in the bucket?

Before consider the question of 3 ounces, let us think about how we might proceed in a different case. Can we put exactly 6 ounces in the bucket? Yes, we can. We first fill the 17-ounce glass with water from the river. We dump all of this water into the bucket. The bucket then has 17 ounces of water in it. Then we use the 11-ounce glass to scoop water out of the bucket. When the 11-ounce glass is full, how much remains in the bucket? There are 6 ounces of water remaining.

Here is another question. Can we put exactly 5 ounces of water in the bucket? Yes, we can. Here is a method. Empty the bucket. Fill the 11-ounce glass from the river, and pour the water into the bucket. Do this again, so that there are 22 ounces of water in the bucket. Now take the 17-ounce glass and scoop water out of the bucket. When 17 ounces have been taken away from 22 ounces, there are 5 ounces remaining in the bucket.

Before continuing, attempt now to solve the puzzle of putting exactly 3 ounces into the bucket using the 11 and 17-ounce glasses.

As you think about this, you might consider doing things like pouring from one glass into the other glass, rather than into the bucket. This is fine. In the end, it does not change the possibilities if we simply assume that water is always poured into the bucket or scooped out of it.

13.1.2 *Weighing Objects*

Now we will consider a different kind of combination puzzle. Suppose that we have many weights that weigh exactly 14 pounds, and many weights that weigh exactly 17 pounds. We also have a balance. This balance is level when the weight on one side is the same as the weight on the other side. It does not have any direct kind of measurement apparatus, it simply says whether the weights on both sides are equal. Finally, we have an object of unknown weight. We want to know whether or not this object weighs exactly 5 pounds. How can we do this?

Let us begin with an easier question. Can we determine whether or not the object weighs exactly 3 pounds? Yes, we can. Place the object and one of the 14-pound weights on one side of the balance. Place a 17-pound weight on the other side of the balance. If the balance is level, the object must weight exactly 3 pounds.

Here is another question. Can we determine whether or not the object weighs exactly 11 pounds? Yes, we can. Place the object and a 17-pound weight on one side of the balance, and place two 14-pound weights on the other side. If the balance is level, the object must weigh exactly 11 pounds.

Before continuing, attempt to solve the problem of determining whether an object weighs exactly 5 pounds using the items given.

Note that it is never useful to put the same weights, say the 14-pound ones, on both sides of the balance. This is because each cancels the effect of the other. Thus, you should consider putting 14-pound weights on one side, 17-pound weights on another side, and then putting the object on one side or the other.

13.1.3 *Impossible Cases*

Each of the puzzles you have considered so far had a solution. Here are two cases where solution is impossible. The way that we see this is by thinking mathematically, using greatest common divisors.

Suppose you have an 6-ounce glass, a 10-ounce glass, a large bucket, and a water source. Can you put exactly 5 ounces of water into the bucket? Take some time to try a few cases. In the end, there will be no way to do it.

Suppose you have many 12-pound weights, many 16-pound weights, and a balance. Can you use these to determine whether an object weighs exactly 6 pounds? Take some time to try a few cases. In the end, it is impossible. You should still explore it, though.

13.2 *A Classification*

In order to treat the preceding puzzles in a mathematical fashion, it is necessary to abstract from some of the particular irrelevant features. Here is the common generalization. We have two numbers, and we wish to consider all the ways that they can be combined. We take some multiple of one of the numbers, and we add or subtract a multiple of the other number. In the case of the water problems, the two numbers are the volumes of the glasses. In the case of the weighing problems, the two numbers are the two weights that are available.

Definition 79. *A sum or difference of multiples of two given numbers is said to be an integral linear combination of those numbers.*

"Integral" refers to the fact that we only use the quantities a whole number of times, not fractionally. "Linear" refers to the fact that the condition we consider involves only the numbers themselves, and not their squares, cubes, or other such quantities.

Here are some examples of how we use this phrase. The number 5 is an integral linear combination of 11 and 17, since it is the same as $(2 \times 11) - 17$. The number 39 is an integral linear combination of 11 and 17, since it is the same as $(2 \times 11) + 17$. The number 3 is an integral linear combination of 14 and 17, since it is the same as $17 - 14$. The number 45 is an integral linear combination of 17 and 14, since it is the same as $17 + (2 \times 14)$.

Definition 80. *A linear Diophantine equation is a problem of the following kind: given one number and two other numbers, determine whether the first is an integral linear combination of the other two.*

Return to the first example of this chapter, where we wished to obtain exactly 3 ounces of water using 11-ounce and 17-ounce glasses. We can reformulate this now using the new terms. This puzzle is a linear Diophantine equation, asking whether 3 is an integral linear combination of 17 and 11.

We add with one glass a certain number of times, and we remove with the other glass a certain number of times. If we ever did both with the same glass, those steps could be removed, since the second cancels the effect of the first.

We also saw some impossible cases. We can use the new terminology there, too. We say that the number 5 is not an integral linear combination of 6 and 10.

You might have found an explanation for this yourself, or simply accepted the claim that was made earlier.

Definition 81. *A solution of a linear Diophantine equation is a specific linear combination that is equal to the given number.*

We can say that $17(1) - 11(1)$ is a solution of the linear Diophantine equation that asks us if 6 is an integral linear combination of 11 and 17. Similarly, $11(2) - 17(1)$ is a solution of the linear Diophantine equation that asks us if 5 is an integral linear combination of 11 and 17.

A linear Diophantine equation can have more than one solution. You can check that $17(10) - 11(15)$ is also equal to 5. This is a more complicated solution, thinking in physical terms. We add 17 ounces ten times, and then remove 11 ounces fifteen times.

In fact, if there is one solution, there are infinitely many.

If a linear Diophantine equation has a solution, we can conclude something about the relation between a greatest common divisor and another number. More specifically, we have the following proposition.

Proposition 82. *Suppose that a first number is an integral linear combination of two other numbers. Then the greatest common divisor of those two numbers divides the first number.*

Proof. Let c be the first number, and let a and b be the numbers whose integral linear combination is c. This means that there are numbers x and y so that $ax + by$ or $ax - by$ is the same as c. Since we wish to show that the greatest common divisor of a and b divides c, it will suffice to show that it divides $ax + by$ and $ax - by$.

Because we work with natural numbers, and not integers, we must enumerate the cases of addition and subtraction separately. If working with integers, it is possible to consider only a single case, since the subtraction can be "absorbed" into the sign of the integer y. If you look in most modern books, you will find linear Diophantine equations involving integers.

The greatest common divisor of a and b divides a and thus divides all multiples of a, including ax. The greatest common divisor of a and b divides b and thus divides all multiples of b, including by. The sum and difference of two numbers divisible by a given number are each divisible by that number. Thus, the greatest common divisor of a and b divides c. □

Was the case $by - ax$ omitted? Or could we say "without loss of generality…?"

The proposition just given provides what is called a "necessary" condition for one number to be an integral linear combination of others. In every case that such a combination exists, the divisibility statement holds. This allows us to consider the impossible puzzles from the previous section in a more precise manner. Rather than simply saying "I never found a way to do it," we can go further and say "No one could ever find a way to do it."

First, consider the 6 and 10-ounce glasses and the goal of measuring 5 ounces. Adding water and then taking it away using these glasses amounts to a linear combination of the numbers 6 and 10. Thus, we wish to know whether 5 is an integral linear combination of 6 and 10. The proposition assures us that this is not so. If 5 were an integral linear combination of 6 and 10, it would be divisible by 2, their greatest common divisor. But 5 is not divisible by 2, and so we know that no such combination exists.

You might be familiar with the term "contrapositive." In a sense we are using the contrapositive of the thing we have shown. Supposing that the divisibility condition is not satisfied, it follows that the one number is not an integral linear combination of the other two.

Now consider the case of 12 and 16-pound weights and the goal of determining an object weight of 6 pounds. We think of the weights being placed on the side opposite the unknown object as counting positively, and then the weights on the same side of the object as taking weight away. We thus want to know whether 6 is an integral linear combination of the numbers 12 and 16. It is not, since their greatest common divisor is 4, which does not divide 6. Thus, we know that this puzzle also cannot be solved.

13.2.1 *Sufficiency*

A question still remains about integral linear combinations. We can conclude impossibility by reasoning about divisibility, but what about possibility? Can we use a divisibility criterion to conclude that one number is in fact an integral linear combination of two others? We can, as it turns out. Not only is the divisibility relation a necessary condition, it is also a sufficient condition. This means that we do not need to find a specific combination in order to be convinced that one exists. Instead, we can know that one exists in a more abstract manner.

Proposition 83. *Suppose that the greatest common divisor of two numbers divides a third number. Then that third number is an integral linear combination of the first two.*

Proof. *See exercises.* □

The proof of the proposition is best understood through some concrete examples, given in the exercises. We can sketch the idea here. The Euclidean algorithm gives us a way to uncover the greatest common divisor of two numbers. By working backwards through the various steps, we can show that the greatest common divisor is an integral linear combination of the two numbers. Then, by multiplying suitably, any multiple of the greatest common divisor is also an integral linear combination of those numbers.

We collect both of the propositions about divisibility and integral linear combinations into a single theorem.

Theorem 84. *One number is an integral linear combination of two other numbers if and only if it is divisible by the greatest common divisor of those numbers.*

The phrase "if and only if" is used when we are able to show that two categories of things are exactly the same. Consider the following statements in Euclidean geometry. The statement "A shape is a square if and only if it is a rectangle" is false. There are rectangles that are not squares, though all squares are rectangles. On the other hand, the statement "A triangle is equiangular if and only if it is equilateral" is true. We showed that the base angles opposite two equal sides are necessarily equal. This is Proposition 14 (I.5). That sides opposite equal angles are the same is Proposition 15 (I.6).

13.3 *Primality and Divisibility*

We defined prime numbers using non-divisibility. Those numbers are said to be prime that are divisible by no numbers but themselves and 1. It turns out that we are able to think about prime numbers in another way, too. This way of thinking about them involves considering them dividing other numbers. In modern areas of mathematics this formulation of primality is useful.

If you look up the term "prime ideal" you will find an instance of such a generalization. You will also encounter interesting mathematical terms that might whet your appetite for newer kinds of mathematics. You can also look at the term "irreducible."

Proposition 85 (VII.30)**.** *A prime number dividing the product of two numbers divides one of the numbers.*

This proposition shows that our Theorem about linear Diophantine equations is not just a triviality for imaginary puzzles. It is a useful, highly significant result.

Proof. Let p be a prime, and let m and n be numbers so that p divides mn. Either p does or does not divide m. If it does, there is nothing further to show.

Suppose, on the other hand, that p does not divide m. This means that the greatest common divisor of p and m is 1, since the only divisors of p are p and 1, and we know that p does not divide m. We then know that there is an integral linear combination of p and m which is the same as 1. This means that there are numbers x and y so that $px - my$ is the same as 1. Multiply each of these expressions by the number n. This yields the expressions $pxn - myn$ and n, and they are the same, since they arose from multiplying the same quantity by the same quantity. Consider the number $pxn - myn$. The number pxn is certainly divisible by p. The number myn is also divisibly by p, since it is a multiple of mn, which was presumed to be divisible by p. The difference of two numbers, each divisible by p, is also divisible by p. Thus, $pxn - myn$ is divisible by p, and this number is just n.

Does the order matter? Can you make the omitted case explicit?

We happened to write it as myn but this is the same as mny or ymn or other permutations.

In conclusion, drawing the two cases together, either p divides m, or p divides n. □

Let us see why numbers that are not prime do not satisfy this condition. Consider the number 6, which is not prime. While 6 divides the product of 3 and 4, which is 12, it does not divide either 3 or 4. Thus, we cannot infer that if 6 divides a product it divides one of the factors.

More generally, given a composite number n which is the product of numbers p and q both larger than 1, the number n divides the product pq (which is just n) but divides neither p nor q.

13.4 Exercises

Exercise 1

Return to the proof of Proposition 82. Find all statements about divisibility used there, and state them clearly as independent propositions. Find where those propositions were proved in an earlier chapter, or prove them yourself.

Exercise 2

Find a solution of each linear Diophantine equation. To find the solution, simply find small multiples of each number and consider sums and differences of these.

a.) Express 14 as an integral linear combination of 6 and 10.

What is 6×4?

b.) Express 9 as an integral linear combination of 5 and 2.

c.) Express 1 as an integral linear combination of 7 and 9.

d.) Express 23 as an integral linear combination of 12 and 13.

What is 12×3?

e.) Express 11 as an integral linear combination of 4 and 7.

Exercise 3

Find additional solutions for each of the linear Diophantine equations in the previous exercise.

Exercise 4

Find three different ways of expressing 2 as an integral linear combination of 12 and 14.

Exercise 5

This exercise shows how the Euclidean algorithm can be used in reverse to find solutions of linear Diophantine equations.

We already know, using basic arithmetic, that the greatest common divisor of 16 and 10 is 2. Nonetheless, let us use the Euclidean algorithm to compute this.

$16 - 10$ is 6
$10 - 6$ is 4
$6 - 4$ is 2
4 is a multiple of 2

Now work backwards. We know that 2 and $6 - 4$ are the same. We also know that 4 and $10 - 6$ are the same. Combine these two equivalences by replacing 4 with $10 - 6$. Then we see that 2 is the same as $6 - (10 - 6)$. That can be rewritten $2 \times 6 - 10$. Now use the fact that $16 - 10$ is the same as 6. Substituting, we find that 2 is the same as $2 \times (16 - 10) - 10$. This latter number can be rewritten as $2 \times 16 - 3 \times 10$. Observe that we have solved a linear Diophantine equation. We have expressed 2 as an integral linear combination of 16 and 10.

Now consider some number divisible by 2. For concreteness, consider the number 8, which is 4×2. We know from what was just done that two copies of 16 minus 3 copies of 10 are the same as 2. Multiplying throughout by four, we conclude that eight copies of 16 minus twelve copies of 10 are the same as 8.

Carry out steps as above to express the number 6 as an integral linear combination of 22 and 14.

1. Use the Euclidean algorithm to compute the greatest common divisor of 22 and 14 (which you know is 2).

2. Work backwards through the series of calculations, making substitutions, to express 2 as an integral linear combination of 22 and 14.

3. Compute the products of the relevant numbers and 3 (which is the quotient of 6 by the greatest common divisor). Check that you obtain an integral linear combination of 22 and 14 that is indeed the same as 6.

Exercise 6

Return to the exercise in the previous chapter in which you computed greatest common divisors using the Euclidean algorithm. Use the previous exercise to write that greatest common divisor as an integral linear combination of the given terms.

Exercise 7

Prove this statement. Given four numbers, suppose that the third is an integral linear combination of the first and second, and the fourth is an integral linear combination of the second and third. Then the fourth is an integral linear combination of the first and second.

Here is the beginning of the proof. You should complete it.

Let m, n, r, and s be the numbers. Since r is an integral linear combination of m and n, there are numbers w and x so that $mw + nx$ is r or $mw - nx$ is r. Since s is an integral linear combination of n and s, there are numbers y and z so that $ny + rz$ is s or $ny - rz$ is s ...

Exercise 8

Explain how the previous exercise shows that the Euclidean Algorithm implies that the greatest common divisor of two numbers is an integral linear combination of them. This amounts to giving the proof of Proposition 83, whose proof we omitted.

14 Numbers in Themselves, Revisited

Having become accustomed to giving proofs involving natural numbers, we can now return to some of the special kinds of numbers that we considered earlier. Rather than simply considering specific numbers, we will reason generally about them and explain their properties demonstratively.

14.1 Mathematical Induction

We will begin with an overview of mathematical induction. It is easier to understand what it is through examples. Thus, you should read this section without slowing down too much, going on to see mathematical induction in action in the sections that follow. Then come back and read this section again.

Each time we consider a collection of natural numbers, that collection has a least element. For example, consider the collection of numbers 7, 19, 2, 5. Evidently 2 is the least element of that collection. In addition to writing down specific numbers, we can also consider more abstract ways of defining collections. A second example is this: the collection of all the even numbers whose square is greater than 71. By thinking about this for a bit, we see that 8 is excluded (as are smaller even numbers), but 10 and greater even numbers are included. Thus, the least member of the collection is 10.

It is possible to use this kind of reasoning to arrive at a new method of proof, called mathematical induction. The idea of mathematical induction is this. Suppose that we wish to prove that a statement involving natural numbers is true for all natural numbers. Then it is sufficient to show two simpler things.

1. The statement is true when applied to the first natural number.

2. The statement being true for one natural number implies that it is true also for the successor of that number.

These conditions must be checked in specific cases (i.e., for specific asserted propositions.)

Why are we convinced that this is a good method of proof? Our goal is to show that some statement holds for all natural numbers. Suppose that it did not hold for all numbers. Then there is a collection of numbers for which it does not hold. By the principle we noted earlier, that collection must have a least element. But then the statement held for the predecessor of that number, and so by part 2. of mathematical induction, the statement also holds for the presumed least number.

Slightly more concretely. Suppose that 100 were the first number for which the statement did not hold. Then the statement did hold for 99. But 100 is the successor of 99, and so the statement also holds for 100.

Keep in mind that there are two things to show. You must show that being true for one number implies being true for the successor. You must also show, however, that there is a single number for which the statement is true. The image of a ladder can be helpful. Part 1. is saying that you can get on the ladder. Part 2. is saying that if you are at one rung on the ladder you can move to the next one. If both of these things are true, you can get to any rung on the ladder.

Warning: Sometimes we use induction but start at a different point than the natural number 1. That is fine. It is still a valid method of proof. It is just using a different starting point on the ladder, which proves the statement for each number greater than the starting point.

14.2 Triangular Numbers

We studied triangular numbers in Section 10.2. There, we simply looked at specific examples. Now we can use mathematical induction to prove that the numbers follow the pattern given in the following proposition. In this use of mathematical induction, our starting point will be 2, rather than 1.

Proposition 86. *The triangular number arising from a triangle with a given number of objects on a side is one half the product of that side number and the successor of the side number.*

The product of a number and its successor will always be even, since one of the two numbers will necessarily be even. Therefore, the product has a well-defined half.

Proof. First, consider the case of the side number 2. The successor of 2 is 3, the product of 2 and 3 is 6, and one half of 6 is 3. Therefore, we see that the assertion holds in the case of the triangular number arising from a triangle with side number 2, since that triangle is made from 3 total objects.

Now consider the triangle with side number n. We wish to show that the corresponding triangular number is one half of $n(n+1)$. We presume that the asserted relationship holds for all smaller numbers.

In particular, we assume that the triangle with side number $n - 1$ yields a triangular number which is one half of $(n-1)n$.

The triangle with side number n is obtained from the triangle with side number $n - 1$ by adding a row of n objects. Thus, we know that the triangular number that we seek is the sum of n and one half of $(n-1)n$. The number n is the same as one half of $2n$. Therefore, the sum we seek can be expressed as one half of $(n-1)n + 2n$. This is the same as one half of $n(n+1)$.

There is a small amount of omitted detail. You can try to work it out now, or do so later in an exercise.

The proposition is true for a successor when it is true for the prior number. The proposition, moreover, holds for 2. Thus, it holds for all natural numbers greater than or equal to 2. □

You can check that the statement also holds for the side number 1, although it might seem silly to say that a single object forms a triangular configuration.

14.3 Square Numbers

It is straightforward to describe the square numbers. Given a number, the product of that number with itself gives the number of objects in a square with the number as the side. Something subtler about square numbers is this.

Proposition 87. *Let odd numbers be summed, starting from 1. The sum is the square whose side is the total number of terms added.*

Try to get a feel for the statement by looking at more examples. What is $1+3+5$? What is $1+3+5+7$?

Proof. The sum of the first two odd numbers, $1 + 3$, is 4. This is indeed the square of 2.

Now suppose that the statement is true for $n - 1$. Observe that the first odd number is $2(1) - 1$, the second odd number is $2(2) - 1$, and in general the k-th odd number is $2k - 1$. This means that in considering the sum of the first n odd numbers, we consider the sum $1 + 3 + 5 + \ldots + (2n-1)$. By hypothesis, the sum of all terms but the last is the same as $(n-1)^2$. Thus, the whole sum is $(n-1)^2 + 2n - 1$. This is the same as n^2.

You can confirm the details of the calculation in an exercise.

Since the statement holds for 2 and it holds for all successors, it holds for all numbers. □

14.4 Tetrahedral Numbers

In addition to plane figures we can consider solid figures. The simplest such is a tetrahedron. We make a tetrahedron by stacking triangles. The first, smallest tetrahedron comes from placing a single object on top of a triangle with three objects. So it has four total objects. The second comes by stacking a triangle with three objects on top of a triangle with six objects, and then putting a single object on top, so it has ten total objects. In general, we stack up triangles one at a time, with each successive triangle having an additional object along a side.

Proposition 88. *The tetrahedral number arising from a tetrahedron with a given number of objects on its side is the one sixth of the product of the side number and its two successors.*

Convince yourself that the product of any three consecutive numbers is divisible by 6.

Proof. Consider the tetrahedron with side number 2. This yields the tetrahedral number 4. The successors of 2 are the numbers 3 and 4, the product $2 \times 3 \times 4$ is 24, and one sixth of that product is indeed 4. Thus, the asserted relation holds for side number 2.

Suppose now that the relationship holds for all side numbers less than n. The tetrahedron with side number n is obtained by adding a triangle with side number n to the tetrahedral number of the tetrahedron with side number $n-1$. We thus consider the sum of one half of $n(n+1)$ and one sixth of $(n-1)(n)(n+1)$. This sum is the same as one sixth of $3n(n+1)+(n-1)(n)(n+1)$. Consider the number $3n(n+1)+(n-1)(n)(n+1)$. It is the same as $n(n+1)(n+2)$, and since the numbers are the same, their sixth parts are the same. This is what was to be shown. □

Half of a thing is the same as a sixth of three times the thing.

14.5 Perfect Numbers

Recall that a number is said to be perfect when it is the sum of all of its proper divisors, where "proper divisor" means those natural numbers that divide it and are less than it. The first two perfect numbers are 6 and 28. Here is a way to generate perfect numbers, under the condition that a certain sum is prime.

Theorem 89 (IX.36)**.** *Given a sum of powers of 2, starting from 1, which is equal to a prime, then the product of the sum and the final summand is perfect.*

We do not work much with 0, but 1 is considered a power of 2 since 2^0 is 1, by convention.

Prior to giving the proof, let us make sense of the statement. We have the sum $1+2$ which is equal to 3, a prime. The product of this sum with 2, the last summand, is 6, the first perfect number. Now consider the sum $1+2+4$, which is 7. This is prime. The product of the sum with 4, the final summand, is 28, which is perfect. The next case to consider involves the sum $1+2+4+8$. This sum is 15, which is not prime, and thus the proposition does not apply.

Proof. Let the sum be p, and let the final summand be 2^k. We must show that $p2^k$ is perfect. We know that the only numbers dividing $p2^k$ are 1, p, powers of 2 up to 2^k, and numbers of the form $p2^j$ with $j < k$. Let us separate these numbers into two classes. One class consists of those that are simply multiples of 2, while the second class consists of those that are divisible by p.

Could you prove that we know this? Consider Proposition 85 (VII.30).

The numbers of the first class are these.

$$1, 2, 4, \dots, 2^k$$

The sum of all these number is p.

Go back the beginning. How did we define p?

The numbers of the second class are these.

$$p, 2p, 4p, \ldots, 2^{k-1}p$$

The sum of the numbers in the second class is $p(2^k - 1)$.

Consider $p + 2p + 4p + \ldots$ in the equivalent form $p(1 + 2 + 4 + \ldots)$ and then use something that you know, from an exercise, about summing powers of 2.

Combining the first and second sums, which together contain all the proper divisors, we find that the sum of all proper divisors is

$$p + p(2^k - 1)$$

which is the same as $p2^k$. Therefore, the sum of the proper divisors of the number is the number itself. □

14.6 Exercises

Exercise 1

In an earlier exercise you made a conjecture about sums of the form $1 + 2 + 4 + 8 + \ldots$; use mathematical induction to prove your conjecture.

Exercise 2

Find the triangular number corresponding to a triangle with 17 objects on a side.

Exercise 3

Show that the sum of all numbers up to a fixed number is the same as the triangular number with that fixed number as its side. Use a figure.

Exercise 4

Write out many numbers in a row. Circle 1, cross off a single number (i.e., 2), circle the next number (i.e., 3). Cross off 2 numbers, circle the next number. Cross off 3 numbers, circle the next number. Keep going for a while. What numbers are circled?

Exercise 5

The number 25 is a square number, being 5^2. It is also the sum of two square numbers, namely 3^2 and 4^2. The collection of numbers $(3, 4, 5)$ can be called a Pythagorean triple for the following reason. Let some segment be set out, and produce a right triangle in this form: one of the legs of the right triangle is 3 copies of the segment, and the other leg of the right triangle is 4 copies of the segment. The hypotenuse will be exactly 5 copies of the segment, by the Pythagorean theorem.

It is interesting to try to find other such collections. One easy way to do this is to consider the same multiple of each term, for example doubling each one in the example to get $(6, 8, 10)$. The more interesting thing to do is to try to find triples that are the smallest of their kind, that do not arise from taking multiples of some smaller collection. That sort of thing is called a *primitive* Pythagorean triple. In such a triple the greatest common divisor of any two of the numbers is 1. So both $(3, 4, 5)$ and $(6, 8, 10)$ are Pythagorean triples, but only the first is a primitive Pythagorean triple.

The following procedure corresponds to a lemma in Euclid's Book X.

Task: Produce a primitive Pythagorean triple.

Procedure:

1. Pick two numbers that are relatively prime and so that one of them is even. Call the larger one j, and the smaller one k.

2. Compute the number $j^2 - k^2$.

3. Compute the number $2jk$.

4. Compute the number $j^2 + k^2$.

5. Check that the three numbers you have just produced form a primitive Pythagorean triple.

Exercise 6

Use the numbers 2 and 1 in the procedure of the preceding exercise to obtain the smallest Pythagorean triple.

Exercise 7

Compute the squares of $j^2 - k^2$, $2jk$, and $j^2 + k^2$ symbolically to confirm that these numbers yield a Pythagorean triple.

Exercise 8

For this exercise, proceed as follows. Given a complicated expression, carry out multiplication symbolically, using the fact that multiplication "distributes" over addition. To give an example, the number $(n-1)n$ is the same as the number $n^2 - n$. By using this property repeatedly with each of the numbers given, and combining similar terms, you can reduce them to the same form. Once they are in the same form you can see that they are indeed equal. Do not try to prove these statements using mathematical induction. Simply proceed directly via symbolic operations. These are small pieces of the reasoning used in larger proofs.

a.) Given a number n, show that the number $(n-1)n+2n$ is the same as the number $n(n+1)$.

This was used in the proof of Proposition 86.

b.) Given a number n, show that the number $(n-1)^2+2n-1$ is the same as the number n^2.

This was used in the proof of Proposition 87.

c.) Given a number n, show that the number

$$3n(n+1)+(n-1)(n)(n+1)$$

is the same as the number $n(n+1)(n+2)$.

This was used in the proof of Proposition 88.

Exercise 9

Find a number that is both triangular and tetrahedral. Can you find a second one?

Exercise 10

Explore square pyramidal numbers, built by stacking up square of ever greater size, and putting a single object on top. Note that the one of them is 5, with a square of 4 as base and then a single object on the top. Can you find more of these numbers? Can you conjecture the pattern by which they arise, and then prove your conjecture?

Exercise 11

Verify that no odd number less than 100 is perfect. Note that you do not need to check prime numbers, since they only have a single proper divisor, namely 1, and hence are much greater than the sums of their proper divisors.

If, one day, you find an odd perfect number—it would need to be very large, since people have checked all the small ones—you will do something that no one has ever done before.

Exercise 12

Verify that the sum $p+2p+4p+\ldots+2^{k-1}p$ is the same as the number $p(2^k-1)$.

Exercise 13

Find as many perfect numbers as you can, using Theorem 89 (IX.36). Do the arithmetic by hand.

Exercise 14

Show that each power of 2 is greater than the sum of its proper divisors, and so is not perfect.

Exercise 15

Conjecture a way to express the sum, starting from 1, of powers of a number. The expression will involve the next power after the largest power in the sum, and some other simple terms. It is good to proceed gradually. Begin with powers of 3.

We have done this for the specific number 2. Here you are to extend that result to other numbers.

$$1+3$$

$$1 + 3 + 9$$

$$1 + 3 + 9 + 27$$

$$1 + 3 + 9 + 27 + 81$$

Once you have guessed a pattern for the number 3, consider the number 4.

$$1 + 4$$

$$1 + 4 + 16$$

$$1 + 4 + 16 + 64$$

Having seen the pattern for 3 and 4, try to make a guess about how it works in general. Having made that guess, see whether it is true for a sum of powers of 7.

$$1 + 7 + 49$$

Exercise 16

Having given, in the previous exercise, a conjecture about the sum of powers of a number, use mathematical induction to prove that your conjectural sum is in fact the sum.

Exercise 17

A pair of numbers is said to be *amicable* if the second number is the sum of the proper divisors of the first, and the first number is the sum of the proper divisors of the second. People have discovered many such pairs, although there are few that involve small numbers. There is one pair of amicable numbers in which each number of the pair is between 200 and 300. Find this pair.

Exercise 18

Cube numbers can be produced as sums of odd numbers. The first cube (we call 1 the first cube) is the first odd number. The second cube is the sum of the next two odd numbers (i.e., the second and third odd numbers). The third cube is the sum of the next three odd numbers (i.e., the fourth, fifth, and sixth). This pattern continues. Prove that this is so. Given a number n, compute these things.

1. The sum of the first $1 + 2 + 3 + \ldots + n$ odd numbers.
2. The sum of the first $1 + 2 + 3 + \ldots + (n - 1)$ odd numbers.
3. The difference between the sum computed in part 1. and the sum in part 2. The claim is that this is the desired cube.

Observe that you need to use two different things that we have studied already. One of them is the sum $1 + 2 + 3 + \ldots + n$. Call that sum S. The second thing you need is to find the sum of the first S odd numbers.

Work with specific small numbers at first to become familiar with what is being asserted.

Exercise 19

Whether or not you were able to complete the proof, use the proposition of the previous exercise to write 12^3 as a sum of twelve consecutive odd numbers. To find the starting odd number, you must progress through

$$1 + 2 + 3 + 4 + \ldots + 11$$

odd numbers. You know how to compute that sum. It is a triangular number. Use the fact that the j-th odd number is $2j - 1$.

Exercise 20

This exercise presents a method for generating approximations of the ratio of the diagonal of a square and its side. Here are three increasingly accurate approximations of that ratio.

$$1 : 1 \qquad 3 : 2 \qquad 7 : 5$$

These approximations are related. We can refer to one of the ratios symbolically as $a : b$. The next one has the form $a + 2b : a + b$. Check that this is so in each case above. Applying that rule to $1 : 1$ gives $3 : 2$, and applying the rule to $3 : 2$ gives $7 : 5$.

Applying the rule to the ratio $7 : 5$, we obtain $17 : 12$. This process can be continued indefinitely.

By the Pythagorean theorem, we know that the square on the diagonal is twice the square on the side. Consider the ratios formed by the squares of the terms in the preceding ratios.

$$1^2 : 1^2 \qquad 3^2 : 2^2 \qquad 7^2 : 5^2 \qquad 17^2 : 12^2$$

In each case, we obtain a ratio that is almost the ratio $2 : 1$. In some cases ($3^2 : 2^2$ and $17^2 : 12^2$), we find that the greater term is one more than twice the lesser. In other cases ($1^2 : 1^2$ and $7^2 : 5^2$) we find that the greater term is one less than twice the lesser. Our goal is to prove that this pattern continues.

1. Suppose that $a : b$ is one of the approximations that we have obtained by the rule given above.

2. The next approximation is $a + 2b : a + b$.

3. Compute the square of the first term $(a + 2b)$, confirming that it is $a^2 + 4ab + 4b^2$.

4. Compute the square of the second term $(a + b)$, confirming that it is $a^2 + 2ab + b^2$.

5. Suppose in the approximation $a : b$ that a^2 is the same as $2b^2 - 1$ (which is like the $7 : 5$ example above, the square of the first term is one less than twice the square of the second). Show that $(a + 2b)^2$ is one more than twice $(a + b)^2$.

6. Suppose, on the other hand, that a^2 is the same as $2b^2 + 1$ (which is like the $3 : 2$ example above, the square of the first term is one more than twice the square of the second). Show that $(a + 2b)^2$ is one less than twice $(a + b)^2$.

7. Conclude, using mathematical induction and the two previous steps, that every ratio $a : b$ generated by the rule will satisfy the condition that a^2 and $2b^2$ differ by 1.

Exercise 21

Explore the sequence of ratios that come from taking $a : b$ and producing the ratio $a + 3b : a + b$. What relationship seems to hold between the squares of the terms? In this case, you will sometimes obtain ratios whose terms are not relatively prime. You should first reduce them, expressing the ratio using numbers that are relatively prime. How much of the reasoning in the previous exercise can you use here? What about producing ratios in the form $a + nb : a + b$ for another natural number n?

15 Relations Between Numbers

Numbers admit of comparison (greater than and less than) and of repetition (multiplication, which is repeated addition). Thus, they are objects that fall under the earlier notion of ratio, just as geometric objects do. We can do more, though, than simply say whether or not two ratios of numbers are the same. We will begin by discussing a broader classification of ratios.

15.1 Coarse Classification of Ratios of Numbers

The first, simplest way that numbers can be related is simply as equal. In this there is nothing more to say.

The next way that numbers can be related is by divisibility. Given two unequal numbers, it is possible that the larger is a multiple of the other. The ratio $6:2$, for example, is a multiple ratio.

Two numbers that are neither equal nor related by divisibility can have to each other a ratio called superparticular.

Definition 90. *Let a pair of distinct numbers be given in which the larger is less than twice the smaller. When the difference between the large number and the small number divides the small number, we say that the numbers have a superparticular ratio.*

Here are some examples of superparticular ratios.

$$4:3 \qquad 6:5 \qquad 9:8 \qquad 24:21 \qquad 35:30$$

Observe that the difference in each of the last two cases is not 1, but that each difference does divide the respective smaller term.

It is important to note that $4:3$ and $6:5$ are not the same ratio. The first is in fact greater than the second, since 3 sixes do not exceed or equal 4 fives. Nonetheless in this coarser classification of ratio, we say that each of those ratios is superparticular.

A fourth way that numbers can be related is called "superpartient." Let two unequal numbers be given, with the larger less than twice the smaller, and which are not in a superparticular ratio. Then the two numbers are said to have a superpartient ratio.

There are additional ways in which numbers can be related (such as "multiple superparticular," an example of which is $7 : 3$) but we will not consider these now.

15.2 *The Rule of Adrastus*

Here is a procedure, sometimes called the rule of Adrastus, to generate all ratios of numbers. This procedure follows the order of the classification given above. We begin with equality, then pass to multiple ratio, then to superparticular ratio, and then to superpartient ratio.

We will illustrate it in a number of examples. You will formally explain various properties in the exercises.

Set out in an array three 1s. Each term is the same as the others.

$$1 \quad 1 \quad 1$$

Produce a new array of three terms. The three new terms are determined by the previous terms in this way; they are the first number, the sum of the first two numbers, and the sum of the first, the third, and twice the second. This gives the following collection.

You might find a symbolic explanation more helpful than the verbal one. If we start with the numbers

$a \quad b \quad c$

the new numbers we get are

$a \quad a+b \quad a+2b+c.$

$$1 \quad 2 \quad 4$$

Repeat the same procedure, placing the first, the sum of the first two, and the sum of the first, the third, and twice the second into an array.

$$1 \quad 3 \quad 9$$

By continuing in this way we obtain every multiple ratio. You will prove this in an exercise.

Now choose one of these triples in which there stands a given ratio. We choose the simplest one, corresponding to the ratio of the double. Reverse the order of the terms.

$$4 \quad 2 \quad 1$$

Now proceed as before, setting out the first, the sum of the first two, and the sum of the first, third, and twice the second.

$$4 \quad 6 \quad 9$$

This gives us two ratios that are the same, $4 : 6$ and $6 : 9$. These ratios are superparticular, and are the same as the ratio of 2 and 3.

Reverse that sequence to obtain this one.

$$9 \quad 6 \quad 4$$

Now apply the procedure again to obtain this one.

$$9 \quad 15 \quad 25$$

We obtain two forms of the ratio $3 : 5$. We see that this is a superpartient ratio, one in which the difference of the terms is twice their greatest common divisor.

15.3 Three Means

Given two natural numbers, there are various conditions under which it is reasonable to refer to an intermediate number as a *mean*.

15.3.1 Arithmetic Mean

A first form of an intermediate, or mean, between two numbers is the arithmetic mean.

When we speak about the subject "arithmetic" (the discipline that involves addition, subtraction, multiplication, etc) we pronounce the word a-RITH-meh-tic, with stress on the second syllable. That is the way the word is said when it is used as a noun, and is almost certainly how you learned to pronounce it. That is how Part II of this book should be called.

In this specific case of studying means, we are using the word "arithmetic" as an adjective rather than as a noun. It modifies the word "mean." It is typical in the United States for mathematicians to pronounce this adjective as EH-rith-MEH-tic, with stress on the initial and third syllables, and with the initial vowel becoming less clearly pronounced, something like the first vowel sound in the word "feather." The stresses are like in the word "geometric."

Definition 91. *A number is said to be the arithmetic mean of two numbers when the difference between the mean and one of the numbers is the same as the difference between the mean and the other number.*

Here are some examples.

- The number 8 is the arithmetic mean of the numbers 6 and 10, because the difference between 6 and 8 is the same as the difference between 8 and 10.
- The number 21 is the arithmetic mean of the numbers 11 and 31, because the difference between 11 and 21 is the same as the difference between 21 and 31.

Proposition 92. *Numbers have an arithmetic mean if and only if they have the same parity.*

The word "parity" refers to evenness and oddness.

Proof. First, we show that numbers having an arithmetic mean have the same parity. Let two numbers s and t be given, with opposite parity. Suppose that a number m were an arithmetic mean. Let the difference between m and s be k. Since m is presumed to be the arithmetic mean, the difference between m and t is also k. Thus, the difference between s and t is $2k$, which is even. The difference of numbers of opposite parity, though, is odd. Thus, there is no such m.

We prove the contrapositive.

Now we show that two numbers having the same parity have an arithmetic mean. Let s and t be two numbers with the same parity, and let s be the smaller. Then the difference $t - s$ is an even number $2k$. Then the number $s + k$ is the arithmetic mean of t and s, since it differs from both s and t by k. □

15.3.2 Geometric Mean

A second kind of intermediate number involves ratio.

Definition 93. *A number is said to be the geometric mean of two numbers when the ratio of the number to the smaller is the same as the ratio of the larger to the number.*

Here are some examples.

- The number 2 is the geometric mean of the numbers 1 and 4, because the ratio $2 : 1$ is the same as the ratio $4 : 2$.
- The number 6 is the geometric mean of the numbers 4 and 9, because the ratio $6 : 4$ is the same as the ratio $9 : 6$.

Proposition 94. *The square of the geometric mean is the product of the extreme terms.*

The "extreme" terms are the ones that generate the mean.

Proof. An exercise. □

Proposition 95. *Let two numbers be given that have a geometric mean. Given a prime number that divides one of the terms, but not the other, then the prime divides the mean, and the square of the prime divides the term divisible by the prime.*

Proof. Let the two numbers be s and t, and let m be the geometric mean. Then m^2 and st are the same, by Proposition 94. Let p be a prime number dividing s but not t. Since p divides s, it also divides the multiple st of s, and since st is the same as m^2, we see that p divides m^2. Since p divides the product $m \times m$, it must divide one of the factors. This means that p divides m. Thus, p^2 divides m^2, which is st. Since p^2 is relatively prime to t, it must divide s. □

Draw this conclusion using the same reasoning that proved Proposition 85 (VII.30).

15.3.3 Harmonic Mean

A third way to consider an intermediate quantity, or mean, mixes both differences (as in the arithmetic mean) and ratio (as in the geometric mean). This third mean is called the harmonic mean.

Definition 96. *A number is said to be the harmonic mean of two numbers when the ratio of the differences between the mean and the extreme terms is the same as the ratio of the extreme terms.*

Here are some examples.

- The number 4 is the harmonic mean of the numbers 3 and 6. The number 6 is in a double ratio with the number 3. Similarly, the difference of 6 and 4 is in a double ratio with the difference of 4 and 3.
- The number 12 is the harmonic mean of the numbers 10 and 15. The ratio of 10 to 15 is the same as the ratio of 2 to 3. That latter ratio is the ratio of the difference of 12 and 10 and the difference of 12 and 15.

15.3.4 Three Means in a Circle

The three means can be depicted in a single figure, as in Figure 15.1. Establish a segment to be considered as the unit. Let s be one number, and t be another. Let AB be the segment obtained from copying the unit s times. Let AC be the segment obtained from copying the unit t times. Let D be the midpoint of the segment BC, and construct the circle with center D and radius DB as shown. The radius of the circle, DE, is the arithmetic mean of the numbers s and t. More precisely, it is the segment obtained by copying the unit segment arithmetic-mean-many times. The segment AF, perpendicular to diameter BC, is the geometric mean. More precisely, it is the segment obtained by copying the unit segment geometric-mean-many times. Finally, let AG be perpendicular to the radius DF. Then GF is the harmonic mean, or more precisely the unit segment copied harmonic-mean-many times.

We are using hyphenated words here in a way that might seem silly, but there is a good reason to do so. We must be sure to distinguish between numbers and segments. The relationship between the two depends on choosing a unit. It is not an absolute relationship.

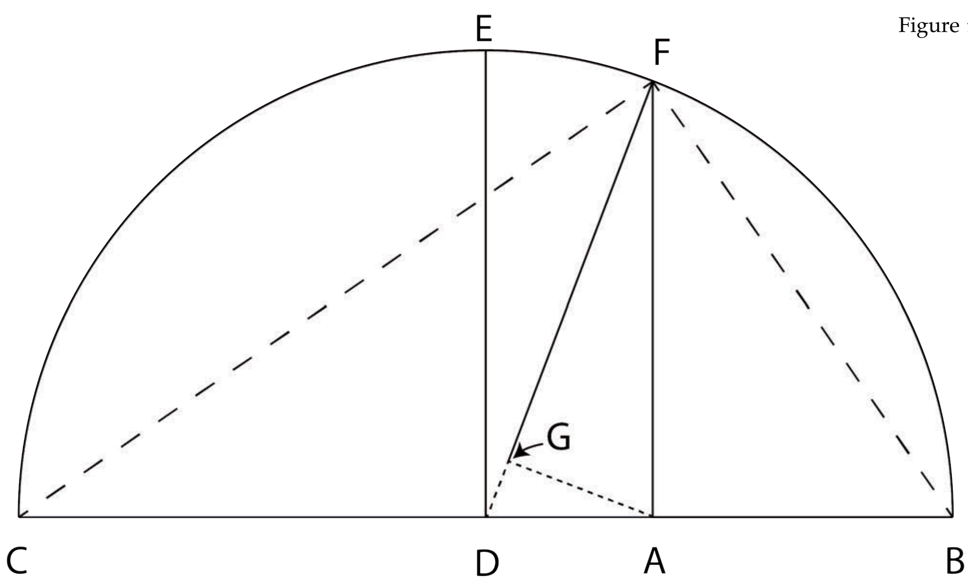


Figure 15.1: Three means

15.4 *Least Terms*

You will show in an exercise that there is no geometric mean between two numbers, one of which is double the other. This places a limit on the means we hope to realize in numbers. Nonetheless we can arrive at a simple collection of four numbers in which, in a certain sense, all three means are realized.

Let us find two numbers having both an arithmetic and a harmonic mean. The harmonic mean involves the ratio of the extreme terms. Let us restrict our attention to the case in which the larger is twice the smaller, since the double is the simplest ratio after identity. Consider some small examples. In the case of the numbers 2 and 4, there is an arithmetic mean, but no harmonic mean, since only the number 3 interposes. In the case of the numbers 3 and 6, there is an arithmetic mean, but no harmonic mean, since the extreme terms have opposite parity. You will check other cases as an exercise.

Now consider these numbers.

$$6 \quad 8 \quad 9 \quad 12$$

We see that the extreme terms 6 and 12 have an arithmetic mean of 9. Furthermore, they have 8 as their harmonic mean. The numbers 6 and 12 do not have a geometric mean. Nonetheless the four numbers above satisfy a relation similar to one satisfied by the geometric mean. The product of the extreme terms is the same as the product of the two intermediate terms. Phrased differently, the ratio of the larger intermediate term to the smaller extreme terms is the same as the ratio of the larger extreme term to the smaller intermediate term.

Recall Proposition 94.

15.5 *Exercises*

Exercise 1

Determine which ratio is multiple, which is superparticular, and which is superpartient. To do this, first check if the ratio is multiple. If it is not, check whether the difference between the terms divides the smaller term.

a.) $6 : 2$

b.) $15 : 13$

c.) $21 : 24$

d.) $100 : 103$

e.) $80 : 64$

f.) 81 : 64

g.) 18 : 16

Exercise 2

Suppose $a : b$ is a ratio of numbers, with a the larger number. Suppose further that $a - b$ divides b. Does $a - b$ also divide a? What does this have to do with the notion of superparticular ratios?

Observe that $a - b$ divides itself. Then use Proposition 64.

Exercise 3

Apply the Rule of Adrastus to each triple. Write down the corresponding ratio in least terms (i.e., terms that are relatively prime).

a.)

$$9 \quad 3 \quad 1$$

b.)

$$16 \quad 4 \quad 1$$

c.)

$$25 \quad 5 \quad 1$$

Exercise 4

Apply the Rule of Adrastus to each triple. Write down the corresponding ratio in least terms (i.e., in terms that are relatively prime).

a.)

$$9 \quad 15 \quad 25$$

b.)

$$9 \quad 12 \quad 16$$

Exercise 5

Show that every multiple ratio arises from the Rule of Adrastus. Do this in the following way. The multiple ratios are of the form $1 : n$, and these correspond to our triples

$$1 \quad n \quad n^2$$

with n some specific natural number. Is the ratio $1 : 2$ something that arises from the Rule of Adrastus? Yes, it is, as we showed above. Now suppose that the ratio $1 : n$ has arisen from the rule. Show that when we apply the rule to

$$1 \quad n \quad n^2$$

we obtain an array of three numbers corresponding to the ratio $1 : n + 1$.

Exercise 6

Consider a triple that arises from a multiple ratio,

$$1 \quad n \quad n^2$$

and reverse it as we did above, to obtain this.

$$n^2 \quad n \quad 1$$

Show that the Rule of Adrastus, applied to this array, yields an array that corresponds to a superparticular ratio. Give the corresponding ratio in least terms. By "least terms" is meant terms that are relatively prime. You will see from this that every superparticular ratio arises from the Rule of Adrastus.

Exercise 7

Let

$$a \quad b \quad c$$

be an array corresponding to a superparticular ratio. Show that by applying the Rule of Adrastus to

$$c \quad b \quad a$$

we obtain a superpartient ratio in which the difference of the terms is twice their greatest common divisor.

Exercise 8

Let

$$a \quad b \quad c$$

be an array corresponding to a superpartient ratio in which the difference of the terms is twice their greatest common divisor. Show that the Rule of Adrastus applied to the array

$$c \quad b \quad a$$

yields an array in whose corresponding ratio is superpartient such that the difference in terms is thrice their greatest common divisor.

Exercise 9

Make a generalization of the previous exercises explaining how all superpartient ratios arise from the Rule of Adrastus.

Exercise 10

Produce triples using the Rule of Adrastus that are neither multiple, nor superparticular, nor superpartient. Explore their properties. Conjecture statements about how they arise, and prove your statements.

Exercise 11

Let numbers a, b, c, and d be such that the ratio of a to b is the same as the ratio of c to d. Show that these ratios are also the same as the ratio of $a + c$ to $b + d$.

This is not a statement about the arithmetic of fractions.

Exercise 12

Suppose that m is the geometric mean of numbers a and b. This means that the ratio $m : a$ is the same as the ratio $b : m$. Take m copies of the first term and b copies of the second term in each ratio. Use these to prove Proposition 94.

Exercise 13

Check the pairs

- 4 and 8
- 5 and 10

and see that in neither case is there a harmonic ratio. This, combined with the observations in the chapter, show that the pair 6 and 12 is the least which reveals all means, among pairs having a double ratio.

Exercise 14

Find the least pair of numbers that stand in a triple ratio and have both an arithmetic and harmonic mean. The first such example involves numbers that are quite small.

Exercise 15

Find the three means of the numbers 10 and 40.

Exercise 16

Find the smallest number a so that a and $9a$ possess all three means. Note that the geometric mean is easy; it is $3a$.

Exercise 17

Complete the proof of the proposition.

Proposition 97. *Two numbers, one of which is double the other, do not have a geometric mean.*

The contemporary formulation of this statement, one you will likely hear, is that "the square root of two is irrational."

Proof. Suppose that there were at least one such such pair of numbers, one of which is double the other, and having a geometric mean. Among all such pairs, consider the smallest one (i.e., the first when counting upwards from 1). Let the smaller number in this pair be a and let the mean be m, so that $m : a$ is the same as $2a : m$. Because these ratios are the same, it must be that m^2 is the same as $2a^2$. This shows that 2 divides m^2, and thus 2 divides m. We conclude that m is even, so that there is a number k so that m is $2k$. Then $4k^2$ is the same as m^2 which is the same as $2a^2$, and thus $2k^2$ is the same as a^2.

You can tidy this argument up a bit by referring to Proposition 95. Do you see how?

(Show that a is even. Show that the numbers one-half-of-a and a have a geometric mean. Conclude that the "smallest pair" hypothesis has been contradicted, so that no such pairs exist.)

□

Exercise 18

Given two numbers having all three means, the geometric mean of the harmonic and arithmetic means is the geometric mean of the original two numbers. Show that this is so.

Exercise 19

Make a copy of Figure 15.1, and prove that the three means are present at the stated places in the following way.

1. (Arithmetic) Observe that AD is the difference of the radius with AC as well as with AB.

2. (Geometric) Use the fact that triangle BFC is right, and so triangles FAB and CAF are similar.

3. (Harmonic) Use the statement proved in the previous exercise, along with the fact that the right triangles FGA and FAD are similar, and that DF is a radius.

Exercise 20

Set out two segments. Produce a single figure containing their arithmetic, geometric, and harmonic means.

Part III

Music

Aperiam in psalterio propositionem meam.

16 Sound

Put your hand on your throat and say "Ahhhh." You will feel a vibration in your throat. Vary the pitch, making it higher or lower, and see how it feels.

Sound arises from the regular vibration of some object. When you speak or sing, there is vibration in your throat, your chest, your mouth, and your head. You are able to control this vibration in order to make intelligible sounds. The vibration that leads to sound is especially evident in stringed instruments. In a guitar, violin, or cello, you can see the object vibrate before you. It is the string.

16.1 Characteristics of Sounds

Three features of sounds distinguish them.

1. Volume. Sounds can be loud or quiet. This aspect, the volume, corresponds to the vigor with which the vibrating object moves.

2. Pitch. Sounds can be high or low. This aspect, the pitch, corresponds to the rapidity with which the object vibrates.

3. Timbre. Sounds can be distinguished even when they are similar in volume and pitch. We refer to this distinction with the term timbre. One way to think about timbre is that it is the way that a single sound is composed of many contributing sounds.

This word rhymes with "amber."

16.2 Time

All sound involves the passage of time. Sound arises from regularity in the movement of an object. This regularity is understood as something that persists with the passage of time.

Speech and music are also given shape through time in another way. Silence and sound alternate in time, and endure for varying measures of time, resulting in rhythm. This introductory study omits rhythm. It is important, though, for the proper production of music and speech.

16.3 Consonance and Dissonance

A single sound, with its volume, pitch, and timbre, is somewhat like a segment in geometry. It is simply a single thing. When we have two sounds, like when we have two segments, we can think of their relation to one another.

Let us suppose that we have fixed an instrument, so that the volume and timbre are relatively unchanging. Now we vary only the pitch. Two pitches that are, together, pleasing and uniform are said to be consonant. Two pitches that are instead harsh and unpleasant, that fail to blend, are said to be dissonant.

When we consider two sounds together we can think of them as occurring either in succession or simultaneously. An example of the first case is when a person plays an instrument like a trumpet, which makes a single note at a time. The second case is like a person playing multiple keys on a piano, or two people singing together.

In our study of music we will consider sequential, rather than simultaneous, sounds.

16.4 Exercises

Exercise 1

Sing various pitches, with varying volumes, on various vowel sounds. Feel your throat and chest while doing so. Describe similarities and differences.

Exercise 2

Make a vowel sound.

1. Make the sound two times with the same pitch and timbre, but with a different volume.

2. Make the sound two times with same volume and timbre, but with a different pitch.

3. Make the sound two times with same volume and pitch, but with a different timbre.

Exercise 3

Drop drops of water into a glass of water. When do you hear the drop? Is the drop ever silent? Is the pitch always the same?

Exercise 4

Listen to the sounds of people's voices. Use the terms "pitch," "volume," and "timbre" to distinguish among them.

Exercise 5

Listen to sounds you regularly hear but that are not made by people's voices. This could be a dog barking, a bird singing, a car passing, or something else. Distinguish among these sounds according to pitch, volume, and timbre.

17
The Monochord

17.1 Preparing Your Instrument

Build, borrow, or buy a monochord. What is a monochord? It is a very simple musical instrument, consisting of a single string. A monochord also has some means of stopping the string at an arbitrary pont along its length. With a guitar or a violin, you do this with a finger. By fixing the string at a specific place you alter the length that is vibrating. A guitar has "frets", designated points at which the string can be stopped. A violin, on the other hand, does not have frets. The violin string can be stopped at any point along its length. A monochord is more like a violin than a guitar.

An outline of how to build a rudimentary monochord is given at the end of this chapter.

Carefully produce, on a sheet of paper, a line with the same length as the length of your monochord string. More specifically, the segment should be the same as the portion of the monochord string that moves freely between the two bridges. You might need to connect multiple pieces of paper together to do this.

17.2 Division of Segments

We will divide the segment representing the monochord string into various pieces. We know how to divide it into 2 pieces of the same size. This involves finding the midpoint, which we can do using a compass and straightedge. The following procedure is more general, allowing us to divide a segment into any number of pieces.

Task: Given a segment AB and a number n, divide AB into n equal parts.

This is a slight generalization of a puzzle from the very first chapter of this book. Did you solve it? How does your solution compare to the one given here?

Procedure:

1. At the point A, produce a new segment AC not in line with AB. The segment AC should be relatively short.

2. Be sure that the point C is clearly marked.

3. Extend AC beyond C.
4. Open the compass to radius AC, and with center C mark an arc on the other side of C from A, yielding a point D such that AC and CD are equal.
5. Repeat the previous step until there are n equal copies of AC, all in a single line. Let the end of this line be E.
6. Produce the line from E to B. You might need a yardstick or some other long object. Do it exactly.
7. Produce the line through C and parallel to BE, and extend it to intersect AB.
8. The intersection of AB with the line through C parallel to BE is one part of the division of AB into n parts.

17.3 The Octave

Work with a segment that is the same length as your monochord string. Use a compass and straightedge to divide the segment in half. Make a mark on your monochord corresponding to this half length. Play the open string. Then play half the string by stopping it at the midpoint. This interval, the relationship between the two pitches, is called an octave.

If your compass is too small compared to the string length, you will not be able to find the midpoint directly as you did in Chapter 2. Instead, use similar triangles like in the procedure of Section 17.2.

Listen to the two pitches of the octave. You might say that they are the same. This equality is subtle, though. One is clearly higher than the other, and yet there is also a sense in which they seem to be the same pitch. This is something that you must hear. It is not something easy to explain in words.

Remember that we used the term "same" in multiple senses in our study of geometry. Look back to Proposition 29 for one example.

You might wonder why the octave is called an "octave," which suggests the notion of the number eight. At this point we consider an octave simply as one thing. Later, when we treat the tone, we will see the way in which the octave is composed of eight things.

17.4 The Fifth and the Fourth

17.4.1 Fifth

Work again with the segment that is the length of the whole monochord string. Divide it into three equal parts. Mark these parts on the monochord. Play the open string, then stop the string at one third, and play two of the three parts. The relation of the sound of the open string to this sound is called a fifth.

The whole string and the fifth are not related in the same way as the whole and the octave. Rather than seeming the same, these pitches are more clearly distinct. They are, nonetheless, consonant. They fit together nicely.

Now play the fifth, which is two of the three equal parts, and then play only a single one of the three parts. You will hear an interval that you have heard before. This interval is an octave. We have produced an octave, not above the whole string, but above the fifth. We can see this through the relation of string lengths. The fifth is two parts, and the new note is one. We have earlier called such a relation of sizes by the name "octave." Hear for yourself that the relation of the fifth to the note which is one of three equal parts of the whole is the same as the relation of the whole string to the half (i.e., the octave above the whole). While the pitches are different, absolutely speaking (some are higher and some are lower), the one pair has the same relationship as the other pair does.

17.4.2 Fourth

Play the whole string and then the fifth a number of times, in that order. Then play the fifth followed by the octave a number of times, in that order. You should still have the octave marked. It is at the midpoint of the whole string. The relationship of the fifth above the whole string to the octave above the whole string is known as the fourth. Listen carefully to this interval and hear that it is also consonant, like the fifth. It is smaller than the fifth.

Now make a new division of the string. Divide it into four equal parts, and consider the sound of the whole and the sound of three of the parts. While the pitches themselves are different, the relationship should sound similar to that of the fifth and the octave. Play the two pairs in order to compare them, in the following way.

1. Play the fifth.
2. Shortly thereafter, play the octave.
3. Let the sound diminish.
4. Play the whole.
5. Play three of four parts of the whole.

Your ear should suggest to you that there is something in common in both cases. We will now show that the same ratio underlies both pairs of sounds.

In order to understand the mathematical character of the fourth, we must consider a division of the string into more parts. The fifth above the whole string is two of three equal parts, and the octave is one of two parts. Consider a division of the string into six equal parts. Then the fifth is four of six parts, and the octave is three of six parts. These numbers are below.

$$\underbrace{\underbrace{6 \quad 4}_{\text{fifth}}\,\underbrace{\quad 3}_{\text{fourth}}}_{\text{octave}}$$

Now consider the fifth as the starting point, so those four parts are treated as the whole. Then the fourth above the fifth, which was considered the octave relative to the original whole, is three of these parts. The interval of a fourth corresponds to the ratio of four to three.

Pay attention to the way that musical intervals correspond to pairs of numbers, and to the way that we present those relationships. As was just done above, we will write some intervals like this.

$$\underbrace{6 \quad 4}_{\text{fifth}}$$

The thing that is called a "fifth" is not a single note. It is a relationship between two notes. Do not think of the numbers 6 and 4 as if they were points along a string, and of the fifth as if it were the thing that arises from the portion of the string between those two points. Think instead about the word "parts" always coming after the number, and read something like the expression above as "the relationship of 6 parts and 4 parts is the fifth." The numbers tell how many parts of the string are being played, for each note in the relationship.

Do not confuse our use of numbers with a use of numbers you might have seen when learning to play the violin or guitar. Those numbers are different. We use our numbers to talk about parts of a string, rather than positions of fingers.

17.5 Exercises

Exercise 1

Review Proposition 43 (VI.2), from the study of ratio. Use it to explain why the procedure for the division of a string works.

Exercise 2

Test your accuracy in dividing long segments. Produce a segment and copy it out five times in the same direction, marking each part. Then use the procedure from Section 17.2, which involves an auxiliary line, to divide the long segment into five parts. See how well you do. Challenge yourself by making the original segment fairly long. You can tape many pieces of paper together if needed.

The segment is already divided into five parts by how you constructed it. You are just trying to divide it in five parts in a different way, using the given parts to test your accuracy.

Exercise 3

Play the octave, then play the fifth on your monochord. Retune the monochord so that the former fifth is now the whole (this involves tightening the string). After retuning, use your marking to play a fourth above the whole. Confirm that this pitch is the same as the original octave.

Exercise 4

In the text, we considered the interval of a fifth above a whole, followed by a fourth above the fifth, and saw that this was an octave. Show that a fourth above the whole followed by fifth above the fourth is also an octave, in the following way. Begin with a string divided into two parts. The whole is 2 parts, and the octave is 1 part. We represent this as below.

$$2 \quad 1$$

We wish to introduce a note that is a fourth above the whole. The ratio of a fourth is the ratio $4 : 3$. To introduce this ratio, divide the string into 4 parts (i.e., divide the original parts in half). Then the whole is 4 of the smaller parts, and the octave is 2 of those parts. We can then interpose the fourth.

$$4 \quad 3 \quad 2$$

What is the interval between the fourth and the octave above the whole? Explain.

Exercise 5

Discuss the numbers

$$12 \quad 9 \quad 8 \quad 6$$

that are the least terms in which all three means are realized. What is the musical significance of these numbers? Consider a string divided in 12 parts, and then consider portions of the string.

Return to the final chapter of Part II: Arithmetic to see the three means and their least terms. Recall that the geometric mean of 12 and 6 cannot be expressed in an exact manner with numbers.

17.6 Simple Monochord Plan

This section offers a simple overview of how to build a monochord. It does not give all the details. You should use the materials and tools that are available to you.

The basic components of the monochord, considered abstractly, are these.

Materials:

- A string.
- A rigid object that holds the string.
- A means of suspending the string above the rigid object so that it can vibrate freely.

- A means of putting the string under tension and keeping it taught.
- (optional) A cavity to resonate with the string's sound.

In practice you can use a guitar string and pieces of wood. To tighten the string, you can use an eye bolt or a ukulele tuning machine. The way that you assemble these depends on the exact parts that you choose.

Figure 17.1 depicts a monochord as seen from the side. The long string is suspended above the body by two bridges that are fixed. A third, movable bridge between these two allows for variation in the string length. The string tightening mechanism is not depicted. The gray rectangle beneath indicates the optional resonant cavity. Without such a cavity, a monochord is like a harp, with a bare string vibrating. With such a cavity, a monochord is like a guitar or violin. You do not need to make a resonant box in order to have a functioning monochord. It is sufficient to have a string held under tension by a simple wooden board.

Figure 17.1: Monochord profile

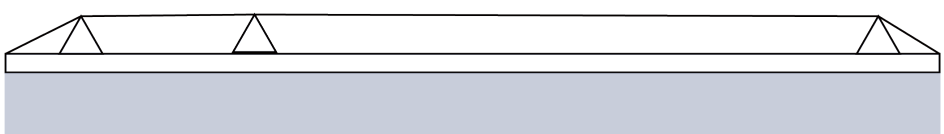
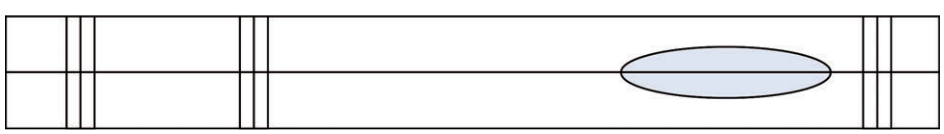


Figure 17.2 shows the monochord as seen from above. The string runs down the middle, the fixed bridges at the end hold it up, and the movable bridge in between allows for playing multiple pitches. The gray ellipse at right indicates an optional hole that could be cut in the top of the monochord if there is a resonant cavity.

Figure 17.2: Monochord from above



To get a clear sound when playing the monochord, apply a small amount of pressure to the string at the movable bridge, just slightly on the side that you do not want to vibrate. This is somewhat like the way that a guitar string is fretted for playing. Be sure, though, that you do not apply too much pressure and thereby alter the pitch.

18
The Tone

We have considered a few ways of dividing the whole string. The first way is to divide the string into two equal parts. This yields the octave. The second way is to divide the string into three equal parts, and to consider the relationship of the sound of two of the parts to the sound of the whole. The third way was to divide the string into four equal parts. We saw that the ensuing relationship between the whole and three parts, called the fourth, also arose between the octave and the fifth.

18.1 The Generation of a Tone

We now introduce the tone, which arises from the divisions that produce the fifth and the fourth. Consider dividing the whole string into twelve parts. We use the number twelve since it is the smallest number divisible by both three and four.

The fifth is two parts, when the whole is three parts. This means that the fifth is eight parts, when the whole is twelve parts. The fourth, which is three parts relative to a whole of four, is then nine of twelve parts of the whole. The ratio of the fourth to the fifth, then, is the ratio of nine to eight. This relationship is called the tone, or sometimes the whole tone, since we will also consider something smaller soon.

Here is the ratio of the fifth, alone.

$$\underbrace{3 \quad 2}_{\text{fifth}}$$

Here is the ratio of the fourth, alone.

$$\underbrace{4 \quad 3}_{\text{fourth}}$$

Here they are together, when the whole has been divided into twelve parts.

$$\underbrace{\underbrace{12 \quad 9}_{\text{fourth}} \underbrace{\quad 8}_{\text{tone}}}_{\text{fifth}}$$

This says that given a note—a pitch or a whole string—moving up by a fourth and then by a tone yields a fifth. You will verify in an exercise that to move up first by a tone and then by a fourth also yields a fifth.

18.2 The Semitone from the Division of a Fourth

We obtain the tone by, in a sense, removing a smaller interval from a larger one. Given a whole and a fifth above it, we consider the whole and the fourth above it, and then obtain the tone through the fourth and the fifth. We proceed in the same way now using the tone, an interval smaller than the fourth, to divide the fourth.

Once we have divided the fourth using the tone we will also have divided the fifth, since the fourth and the fifth differ by a tone.

The relationship of a fourth is found in the ratio of 4 to 3. This is indicated symbolically here.

$$4 \quad 3$$

In order to introduce a tone, which is found in the ratio $9 : 8$, we must introduce a further division of the string. Let us introduce a tone above the whole. Since 4 and 9 have no common factor, we divide the string into 36 parts, 36 being the product of 4 and 9. This leads to the following, which is simply another name for the interval of a fourth.

$$36 \quad 27$$

Verify that this is a superparticular ratio.

This pair of numbers indicates the relationship of a fourth when the whole is divided into 36 parts. To have a tone above the whole, we take 8 of the 9 parts, or equivalently 32 of the 36 parts.

$$36 \quad 32 \quad 27$$

Now let us consider moving down (in pitch) from the fourth by a tone. This means that the 27, which indicates a fourth above the whole, must correspond to 8 parts, in relation to some 9 parts which are to be found. Since 8 and 27 have no common divisor, we divide each of the existing parts into 8 parts, which yields this relationship.

$$288 \quad 256 \quad 216$$

In other words, we suppose that the whole has been divided into 288 tiny parts. In this case, the tone above the whole is 256 parts, and the fourth above the whole is 216 parts.

We now introduce a number corresponding to a tone below the fourth indicated by 216. Since 216 was obtained as 8×27, the quantity we desire is 9×27 which is 243. This gives us a tone below a fourth, when the whole is divided into 288 equal parts.

$$288 \quad 256 \quad 243 \quad 216$$

The intermediate pitches, those corresponding to 256 and 243, do not admit a whole tone between them. You will explore this in an exercise. As a result, the division of the fourth is complete. The fourth can be constructed from a whole tone, an interval corresponding to the ratio 256 : 243, and another whole tone. This is shown below.

$$\underbrace{\underbrace{288 \quad 256}_{\text{tone}} \quad \underbrace{243 \quad 216}_{\text{tone}}}_{\text{fourth}}$$

We use the name "semitone" for the interval, the difference in pitch, arising from the ratio 256 : 243.

18.3 Equivalences

We determined the ratio of the semitone in the previous section by making choices about how to constitute a fourth from smaller pieces. We will now see that this new interval, the semitone, does not depend on these choices. We obtain the same ratio either from moving up twice from the whole string by a tone, or by moving down from the fourth twice by a tone.

First, we move up from the whole. We begin with the ratio of a fourth.

$$4 \quad 3$$

In order to move up from the whole (4 parts) by a tone, we divide the parts by 9.

$$36 \quad 27$$

Proceeding as we did before, we can now introduce a tone.

$$36 \quad 32 \quad 27$$

We wish to introduce a tone above the tone above the whole, but the number 32 is not divisible by 9. Thus, we divide the each of the small parts again into 9, yielding these ratios.

The repetition is not a typo.

$$324 \quad 288 \quad 243$$

Finally, to take 8 of 9 parts, when 288 is the whole, is the same as 32×8 which is 256.

Thus, we have these ratios, with the semitone present once again.

$$\underbrace{324 \underbrace{\quad}_{\text{tone}} 288 \underbrace{\quad}_{\text{tone}} 256 \quad 243}_{\text{fourth}}$$

Finally, we consider the possibility of moving down from the fourth by a tone, twice. Begin with a fourth.

$$4 \quad 3$$

In order for the three of four parts to be 8 parts in a tone, we divide the parts into eight pieces each.

$$32 \quad 24$$

Now we can introduce a tone below the fourth.

$$32 \quad 27 \quad 24$$

We wish again to introduce a tone below the pitch, which is 27 parts when the whole is 32. To do this, we again divide each part into eight further parts. This yields these quantities.

$$256 \quad 216 \quad 192$$

Finally, the tone below the pitch indicated with 216 is 9×27 parts, which is 243.

$$256 \quad 243 \quad 216 \quad 192$$

18.4 Exercises

Exercise 1

Show that the difference between a third part and a fourth part is a twelfth part, directly, using ratio rather than assertions about the arithmetic of fractions. (Hint: Repeat things twelve times.)

Exercise 2

Consider the interval that arises from the tone above the whole and the fifth above the whole. Verify that this interval is a fourth. (Note that this is the opposite way from what was done in the text, where we moved down from the fifth by a tone.)

Exercise 3

Determine the ratio of a tone by finding the difference between the fifth and the fourth, as in the text.

Exercise 4

Determine the ratio of a semitone by dividing a fourth using the tone.

Exercise 5

To move up twice by a fourth is less than an octave. By how much is it less than an octave? Show this directly.

Exercise 6

To move down twice by a fourth is less than an octave. By how much is it less than an octave? Show this directly.

Exercise 7

Show that the semitone is less than a tone by comparing ratios of numbers.

More specifically, given numbers a, b, c, and d, it is easy to compare the ratio $a : b$ to the ratio $c : d$. Simply consider b copies of the first term, and a copies of the second. Then continue.

19 Approximation

We now have a collection of intervals and corresponding ratios. These are collected in the table.

Interval	Ratio
Octave	2 : 1
Fifth	3 : 2
Fourth	4 : 3
Tone	9 : 8
Semitone	256 : 243

It will be helpful for you to review the definition of ratio. Recall also how you can compare two ratios. We will rely on that material here.

19.1 Tone, Semitone, and Comma

The semitone is, as suggested by its name, a part of the tone. It is not, however, exactly half a tone. What is meant by this is that two consecutive semitones are not the same as one tone. We can see this as follows.

$$256 \quad 243$$

The pair above represents one semitone, when the whole has 256 parts. We wish to introduce a semitone above the pitch indicated by 243. To do this, we subdivide each part into 256 parts. This gives these quantities.

One term is 256×256, and the other is 243×256.

$$65536 \quad 62208$$

Finally, another semitone above the pitch indicated by 62208 is given by 243×243.

$$65536 \quad 62208 \quad 59049$$

We must now compare the ratio $65536 : 59049$ to the ratio $9 : 8$. Consider the first number repeated 8 times and the second number repeated 9 times. In the case of the second ratio we obtain the relation of identity, as the number in each case is 72. On the other hand 65536×8 is 524288, and 59049×9 is 531441, which is greater. This means that the ratio 65536 : 59049, the ratio formed from two semitones, is less than 9 : 8, the ratio of a single whole tone.

The numbers 524288 and 531441 differ by a number that is somewhat small in comparison with them. Thus, while their ratio is not the ratio of unity, it is somewhat near the ratio of unity. We can therefore think that a tone is an approximation of two semitones. In ratios, we can say that the ratio 9 : 8 is an approximation of the ratio 65536 : 59049.

Their difference is itself fairly large, but considered relatively it is small.

Let us now consider the semitone alone, the ratio 256 : 243. We wish to find an approximation of this using smaller numbers. More specifically, we will find superparticular ratios that are near to the ratio 256 : 243.

Recall that superparticular ratios are those arising from consecutive natural numbers, like the ratio 4 : 3. More generally, they are ratios in which the difference of terms divides the terms.

The tone is from the ratio 9 : 8. There is no room for any number to interpose, since 9 and 8 are consecutive. Doubling terms yields 18 : 16. We now have the superparticular ratios 18 : 17 and 17 : 16 to consider. Let us compare 18 : 17 to 256 : 243. The product 256×17 is 4532, and the product 243×18 is 4374. We see, then, that 256 : 243 is a ratio less than the ratio 18 : 17.

Consider the next superparticular ratio, 19 : 18. Since 256×18 is 4608, and 243×19 is 4617, we see that 256 : 243 is less than 19 : 18.

Now consider the superparticular ratio 20 : 19. Since 256×19 is 4864, and 243×20 is 4860, we see that 256 : 243 is greater than 20 : 19.

We conclude that the semitone 256 : 243 lies between the superparticular ratios 19 : 18 and 20 : 19.

We can compare the exact semitone (see Exercise 9) to 17 : 16 in an interesting way. The number 17 is the arithmetic mean of 18 and 16. Looking back at Figure 15.1, we see that the geometric mean is less than the arithmetic mean. (The arithmetic mean is the hypotenuse of a right triangle with the geometric mean as a leg.) Thus, the ratio of the semitone must be less than the one that arises from the tone and its arithmetic mean. Do not worry if you do not understand this reasoning. It is not essential for what follows.

19.2 The Octave and the Comma

Six tones are not the same as an octave. An octave consists of a fifth and a fourth. The fifth is three tones and a semitone, and the fourth is two tones and a semitone. The octave is, then, five tones and two semitones, and we saw earlier that two semitones are less than a tone. We will now find the least terms in which to express the relation between six tones and an octave.

Begin by setting out a tone.

9 8

The whole is 9 parts, and 8 parts are a tone above the whole. Continues proceeding upwards. In order to produce a tone above 8 parts, we divide each part again into 9 parts. This gives us the following.

81 72

Then another tone can be introduced.

81 72 64

Continue in this way. Verify for yourself that the numbers are as follows.

729 648 576 512

6561 5832 5184 4608 4096

59049 52488 46656 41472 36864 32768

531441 472392 419904 373248 331776 294912 262144

When the whole has been divided into 531441 parts, six tones above the whole consist of 262144 parts. We now consider the pitch that is one octave below the 262144 parts. To find this we find the number which is twice 262144, since the octave arises from the ratio 2 : 1. Since 262144×2 is 524288, we see that the relation 531441 : 524288 is the difference between the octave and six tones. This is called the comma.

The comma, expressed as a ratio, involves large numbers beyond the range of those with which we have ordinary experience. It is possible to find approximations of the comma using simpler ratios, just as we found approximations for the semitone. Again, we will seek to find consecutive superparticular ratios. Our treatment here will be somewhat brief.

The difference between 531441 and 524288 is 7153. We consider the following multiples of 7153. The first multiple of 7153 greater than 524288 is the product 74×7153, which is 529322. Let us consider the ratio 74 : 73 and compare it to the comma. You can check the computations and see that the ratio 74 : 73 is greater than the comma. Now consider the next superparticular ratio, 75 : 74. Again, you can verify the calculation, which shows that this ratio is less than the comma. We thus have approximations of the comma by simple, superparticular ratios.

You can obtain these multiples by dividing either of the large numbers by 7153 and ignoring remainders.

19.3 Octaves and Fifths

Two fifths are greater than an octave. To see by how much, recall that a fifth is three tones and one semitone, while an octave is five tones

You can also think of this by recalling that an octave is a fourth and a fifth, and a fifth is a fourth and a tone.

and two semitones. Two fifths, then, are six tones and two semitones, so that two fifths exceed the octave by exactly one tone.

We can continue reasoning in this way. Four fifths exceed two octaves by two tones. Six fifths exceed three octaves by three tones. Eight fifths exceed four octaves by four tones. Ten fifths exceed five octaves by five tones. Finally, twelve fifths exceed six octaves by six tones. Recall, though, that six tones themselves exceed an octave, and that the interval by which they are greater is the comma. Thus, we see that twelve fifths exceed seven octaves by a comma. The comma is a small interval, so we can consider seven octaves as an approximation of twelve fifths, or vice versa.

19.4 Thirds

The interval produced by two consecutive tones is called a third. We compute as we have before, dividing the whole into further parts as needed.

$$9 \quad 8$$

$$81 \quad 72$$

$$81 \quad 72 \quad 64$$

Thus, the third is the ratio 81 : 64. This is nearly a superparticular ratio. The ratio 80 : 64 is superparticular, since 16 divides both 80 and 64. In simpler terms, the ratio 80 : 64 is the same as 5 : 4

The third can also be called a major third to distinguish it from the minor third, which is a tone and a semitone.

We arrived at the tone in a natural, simple manner, by considering the difference of the fifth and the fourth. Once the tone is fixed, the interval of two tones is also fixed. It is possible to choose different collections of intervals so that the third will in fact be the ratio 5 : 4, but this leads to other complications.

Observe that 5 : 4, the simple superparticular approximation of our major third, is the first ratio of consecutive numbers that does not occur in our system. The ratio 2 : 1 is an octave, the ratio 3 : 2 is a fifth, and the ratio 4 : 3 is a fourth. Once we have fixed these we immediately obtain the tone, which is 9 : 8.

You should compare this to the reasoning that you have done earlier in geometry and arithmetic. We begin with postulates and definitions that guide our future reasoning. By doing this we impose a constraint on ourselves. These constraints are also, however, sources of knowledge.

19.5 *Impossibilities*

We saw above that twelve fifths do not yield a precise number of octaves. Instead, they differ from seven octaves by a comma. We might wonder whether we might arrive at an octave through a greater number of fifths. This is not so. We can prove it by using things we know from arithmetic.

Proposition 98. *There is no number of fifths that is equal to some number of octaves.*

Proof. Suppose to the contrary that there were numbers m and n so that m fifths were the same as n octaves. By working up from a whole by a fifth, we introduce a division of the whole into three pieces.

$$3 \quad 2$$

$$9 \quad 6 \quad 4$$

Having done this m times, for m fifths, we arrive at these numbers.

$$3^m \quad 2 \times 3^{m-1} \quad 2^2 \times 3^{m-2} \quad \ldots \quad 3 \times 2^{m-1} \quad 2^m$$

Thus, the ratio of m fifths is $3^m : 2^m$.

On the other hand, n octaves proceed as follows.

$$2 \quad 1$$

$$4 \quad 2 \quad 1$$

$$2^n \quad 2^{n-1} \quad 2^{n-2} \quad \ldots \quad 4 \quad 2 \quad 1$$

Thus, the ratio of n octaves is $2^n : 1$.

In order to relate the n octaves and the m fifths, we consider a division of the string into $2^n \times 3^m$ parts. With this whole, the m fifths are the ratio $2^n \times 3^m : 2^n$ and the n octaves are the ratio $2^n \times 3^m : 3^m$.

The assumption that the two are the same means that the same portion of string is 3^m small parts and 2^n small parts. This is impossible, though, since 3^m is odd and 2^n is even, and no number is both even and odd. □

What propositions justify these assertions about parity? What about mathematical induction? Does it belong here? Think about justifying the statements "every power of 3 is odd" and "every power of 2 is even."

The result can be made more general. Suppose that we consider any ratio (of numbers) other than the octave. No number of repetitions of that interval will ever yield an exact number of octaves. An exercise contains an outline of the proof.

We use the name "semitone" to refer to an interval that, when doubled, does not equal a tone. This is reasonable because there is no intermediate within a tone that is expressible using natural numbers.

Proposition 99. *The tone is not divided in two equal intervals by a ratio of numbers.*

Proof. The tone is the ratio $9 : 8$. Consider a string divided into nine equal parts. Then eight of the parts give a tone above the whole. Suppose that there were a musically equal division of the tone using a ratio of numbers. Then there would be a number k into which each of the nine parts is equally divided, and a number a of such parts, so that the ratio $9k : a$ and the ratio $a : 8k$ were the same.

The number a would be the geometric mean of $9k$ and $8k$.

If there is such a division of the nine parts into k smaller parts, there is a least such division. In other words, we consider the smallest k satisfying the condition.

Think back to one of the principles upon which we based mathematical induction. Each collection of natural numbers has a least element.

Observe that a and k cannot both be even, since if they were, their halves would provide smaller terms in which the perfect semitone could be expressed. Thus, our assumption that k is the minimal division of the parts into smaller parts means that a and k are not both even.

The assumption that the number a gives a perfect intermediate dividing the tone means that the ratios $9k : a$ and $a : 8k$ are the same. Taking the first term a times, and the second term $9k$ times, we obtain from the first ratio the ratio of unity. In the second ratio we obtain a^2 and $72k^2$, which must be the same, by the hypothesis that the ratios were the same.

If a^2 is the same as $72k^2$, which is even, then a^2 and hence also a must be even. Since a is even, there is a number b so that a is the same as $2b$, and so a^2 is the same as $4b^2$. Then $4b^2$ is the same as $72k^2$. We see that b^2 is the same as $18k^2$, so that b^2 is even, and so b is as well. This means that there is a number c so that b is the same as $2c$, so b^2 is the same as $4c^2$, and thus $4c^2$ is the same as $18k^2$. This implies that $2c^2$ is the same as $9k^2$, so that 2 divides $9k^2$. Since 2 is prime and does not divide 9 it must divide k^2, which means k^2 is even, and hence k is even as well.

Recall Proposition 85. The number 2 is prime.

This leads to a contradiction of our assumption that k was a minimal further division of the string, since a and k were both shown to be even.

Since there is no minimal further division of the string producing an exact semitone, there is no further division of the string producing an exact semitone.

□

19.6 Exercises

Exercise 1

Compare three tones to the combination of a semitone and a fourth.

Exercise 2

Find the best approximation to two semitones among the ratios 9 : 8, 10 : 9, 11 : 10, and 12 : 11.

Exercise 3

We only considered superparticular ratios when finding approximations of the semitone. Now expand the search. Compare the superpartient ratio 39 : 37 to the semitone.

Exercise 4

Reproduce the calculation of the ratio of the comma. Carry out all calculations by hand. To speed things up, observe that you do not need to keep track of the intermediate terms. So, for example, after introducing two tones you have these numbers.

$$81 \quad 72 \quad 64$$

To move on to the next stage you can ignore the 72. You must remember, though, how many tones you have introduced. One way to do this is to keep intermediate terms but not to carry out computations. As an example, write like this.

$$81 \times 9 \quad 72 \times 9 \quad 64 \times 9 \quad 64 \times 8$$

Then, at the very end, carry out only those computations that are necessary. It is good at times in mathematics to avoid doing unnecessary computations. If you know which ones are unnecessary it means you have a clearer grasp of the essence of the thing being considered.

Exercise 5

Verify that 74 : 73 and 75 : 74 are, respectively, greater than and less than the comma.

Exercise 6

Try to find a general strategy for finding superparticular ratios that approximate arbitrary ratios.

Exercise 7

Compare the comma to the tone.

Exercise 8

Compare the comma to the semitone.

Exercise 9

Use a circle to produce an exact semitone (i.e., as the geometric mean of 9 and 8, see Figure 15.1). Use your monochord to compare the sound of this note to the semitone 256 : 243.

Exercise 10

Copy the proof that there is no exact semitone in numbers.

Exercise 11

Generalize the proof that there is no exact semitone in numbers to the following statement.

Proposition 100. *No superparticular ratio is exactly divisible into two equal intervals by a ratio of numbers.*

Use the fact that a number and its successor (the sort of thing that you use to make a superparticular ratio) are relatively prime.

20
The Diatonic Genus

So far we have developed a mathematical account of certain individual musical intervals using the theory of ratio. Now we will produce a whole system of intervals. This is known as the diatonic genus. The term “diatonic” refers to the fact that we proceed “through,” i.e., by means of, the tone. The term diatonic was used to distinguish this collection of intervals from others. Those other collections were used more frequently in the past, but we will still use the term diatonic, recognizing that we have made a choice to build our construction on the tone.

20.1 The Goal and the Path

We will begin with a monochord having no divisions or markings. Our goal is to produce a system of markings that is intelligible and that allows us to play natural series of pitches on the instrument.

In order to arrive at this goal, we must do a number of things. We must understand the mathematical relations undergirding the intervals. We did this in preceding chapters. Next, we must interpret those relations in a practical manner in order to obtain useful procedures for dividing a physical monochord. Finally, we must put the procedures to work and obtain the desired division.

20.2 A Universal Tetrachord

We know that the fourth is two tones and a semitone. It is divided this way whether or not the semitone is placed between the two tones, or above them, or below them. We will now produce a single diagram that will be useful for producing fourths, divided into tones separated by a semitone. This will let us break an octave into its constituent intervals.

20.2.1 *Producing a Universal Tetrachord*

1. Take a large piece of paper, connecting multiple pieces together if needed. Produce a short segment AB near one edge of the page, perpendicularly to the edge.

2. Mark the point B clearly.

3. Extend AB past B to a point C, so that AC is equal to nine copies of AB.

4. At C, produce a line perpendicular to AC.

5. Mark point D on that line, so that CD is perpendicular to AC and has the same length as AC.

6. Find the midpoint of CD, and find the midpoint of the half of CD that ends in D. Mark this midpoint as E. (Note that this means that CD is four times as long as ED.)

7. At C, and perpendicularly to AC, produce a new line CF on the other side of AC as D so that CF is equal to three copies of AB.

8. Check your progress with Figure 20.1.

9. Draw the line AD.

10. *(Draw this line lightly.)* Draw the line that is through B and parallel to AD. It intersects CD in a the point; call that point G.

11. *(Draw this line lightly.)* Draw the line BE.

12. Draw the line through A and parallel to BE. Let H be the point at which it intersects CD.

13. Check your progress with Figure 20.2.

14. Erase the lines emanating from B, leaving a clear indication of the points E and G at which those lines intersect CD.

15. Produce the lines AE, AG, and AF. They can be extended past CD, and they should be if your monochord string is longer than AC.

Upon completing the steps above, your figure should look like Figure 20.3.

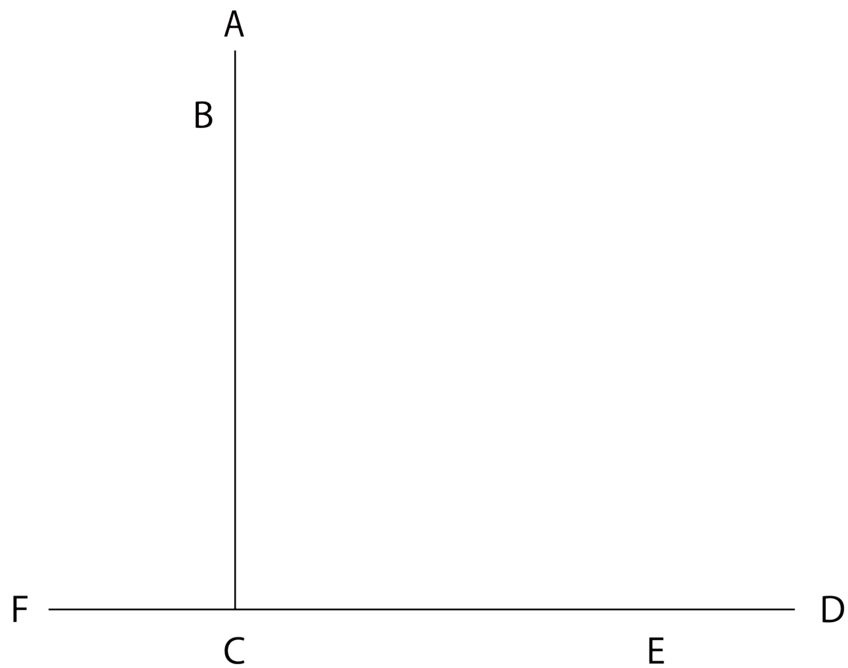


Figure 20.1: Universal tetrachord: initial points

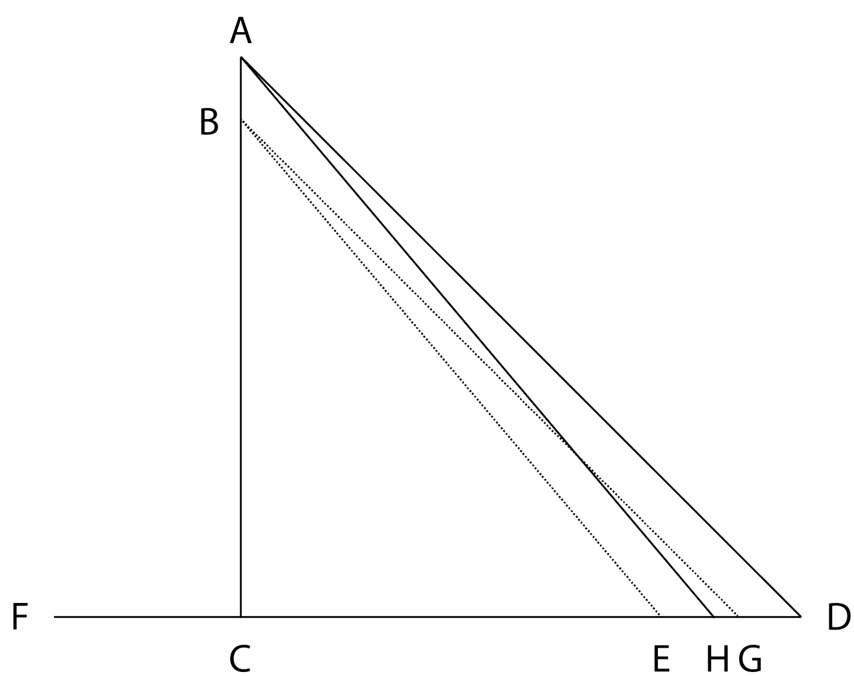


Figure 20.2: Universal tetrachord: auxiliary lines

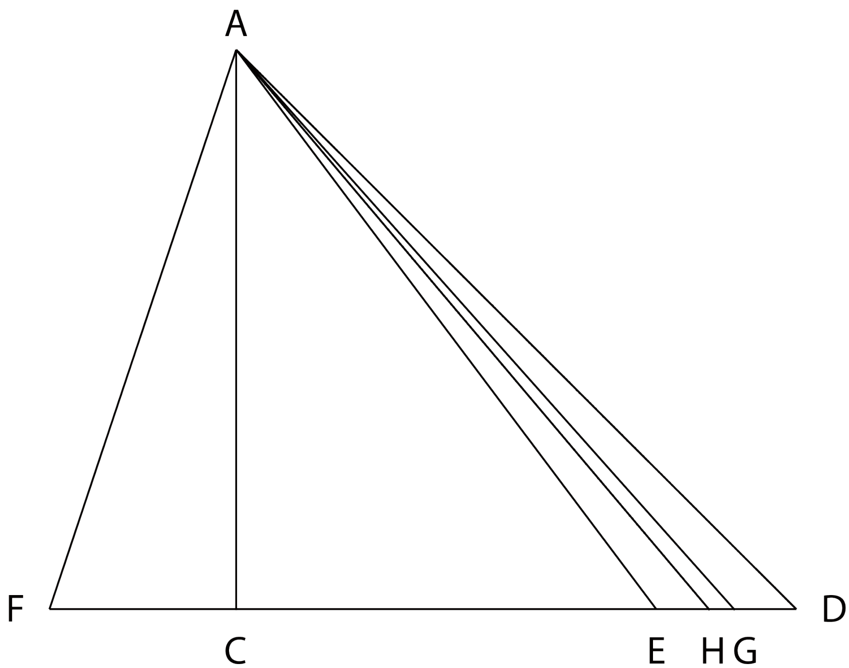


Figure 20.3: Universal tetrachord: complete

20.2.2 *Using a Universal Tetrachord*

The universal tetrachord we have produced allows us to do two things. One is to produce a fourth above a given pitch. This fourth is also divided into a tone, a semitone, and a tone. The second thing that can be done is to produce a fourth below a given pitch.

Consider Figure 20.3. If we take CD to indicate a whole string, then CE is the fourth above the whole. The segment CG is a tone above the whole, and CH is the tone below the fourth. Using properties of similar triangles, we see that we can use such a figure to produce divided fourths for strings of arbitrary length.

Task: Produce a fourth above a given note, and divide the fourth into a tone, a semitone, and a tone.

Procedure:

1. Along the line AC, starting from A, mark out the length of the given string.
2. At the endpoint of that length, produce a line ℓ perpendicular to AC.
3. The distance along ℓ from AC to AE is the fourth above the whole.
4. The distance along ℓ from AC to AG is the tone above the whole.

5. The distance along ℓ from AC to AH is the tone below the fourth above the whole.

We can also work downwards. The next procedure simply produces a fourth below a given note.

Task: Produce a fourth below a given note.

Procedure:

1. Along the line AC, starting from A, mark out the length of the given string.

2. At the endpoint of that length, produce a line ℓ perpendicular to AC.

3. The length of ℓ from AF to AD is the fourth below the given note.

If we wish to divide the fourth below a given note into a tone, semitone, and tone, we can do so by combining the two procedures. First, find the fourth below the given note, using the second procedure. Then divide the fourth above that newly found note, using the first procedure.

20.3 An Octave and More, Marked

Find the length of your entire monochord string. Divide it into four parts, and take three of them. This section of the string, the three parts, will be what we consider as the "whole" and we will refer to it as such. The reason we start with a part of the string is to allow further notes to be added below our foundational note.

1. Divide the whole (remember that this means three quarters of the physical whole) in half, and mark the half. This is an octave above the whole.

2. Mark a subdivided fourth above the whole using the universal tetrachord.

This uses the first procedure of Section 20.2.2.

3. Mark a fourth below the octave.

This uses the second procedure of Section 20.2.2.

4. Subdivide the fourth below the octave.

5. Produce a subdivided fourth above the octave.

6. Produce a fourth below the whole.

This is the real physical whole, if you were careful in the beginning.

7. Subdivide the fourth below the whole.

20.4 Exercises

Exercise 1

Listen closely to hear the difference of the thirds given by 81 : 64 and 5 : 4. We have not yet constructed the latter ratio, but you can do so with the procedure of Section 17.2.

Exercise 2

Produce a subdivision of an octave in the diatonic genus.

Exercise 3

Use similar triangles and ratio to show that our procedure for producing subdivided fourths is correct.

Exercise 4

Given a physical string, it is possible to refer to any portion of it as a whole. This is like establishing a unit of measurement. Briefly discuss the similarities and differences as you produce octaves and fourths below and above the given "whole." What limitation occurs when going in one direction that does not occur in the other? Recall that you are to consider a specific physical string.

Exercise 5

Our universal tetrachord is simple. If you wish, you can supplement it with auxiliary lines between AF and AC to obtain an immediate subdivision of the fourth below a given length. Do this.

Exercise 6

Produce a universal octave. There are various ways you can interpret this.

Exercise 7

We made AC and CD of the same length and perpendicular. It is in fact possible to make a figure that functions just as well as our universal tetrachord if CD is not perpendicular to AC and CD is a different length than AC. Explore this. Carefully explain how you would use such a figure.

Exercise 8

(Requires a large amount of space.) Make a little segment, and copy it in a straight line 288 times, starting from A in your universal tetrachord and going along the line AC, which you might need to extend. At the endpoint of the 288 copies, produce a perpendicular line to AC. Examine the part of that perpendicular between AG and AH. Check whether that part of the line is exactly 13 copies (note

This exercise clarifies why our universal tetrachord matters. We are able to produce ratios of fairly large numbers without carrying out any copying a large number of times. We do this by carefully considering geometric ratios.

that 13 is the difference of 256 and 243) of the little segment that you started with. It is probably not exact. If not, where have errors accumulated?

Exercise 9

Reflect on your experience in the previous exercise. What do you consider to be the best way to construct the ratio 256 : 243?

21
Gregorian Modes

Gregorian chant is an ancient form of music that is still used today. We will use our knowledge of the relation of pitches to study the classification of Gregorian chant by what are called modes.

In music we experience variation and stability, tension and resolution. Beauty in music arises from the artful combination of these elements. Music should be neither sentimental nor lifeless. The modes of Gregorian chant explain, in a structured way, how motion and rest work together in a single musical piece.

It is good to apply ourselves to understand the order that lies beneath things we find appealing. Boethius, the author whose writings have shaped our study of arithmetic and music, says that we should not simply find music pleasing, but instead should discover the structure it has in ratios of pitches.

21.1 Terminology

There are eight modes. They are called by their numbers. In order to understand the distinctions between the modes we need two terms. These terms are *tonic* and *dominant*.

The tonic is the pitch at which a chant concludes. The chant often begins on the tonic as well. The tonic is a kind of natural foundation or resting place for the piece as a whole.

The dominant is a second pitch lying above the tonic at which tension develops. It is itself a kind of secondary foundation for the chant, but seeks resolution in the tonic. When chant is complemented with psalm verses, these verses are chanted almost exclusively on a single pitch. This pitch is the dominant.

21.2 Principles of Classification

In order to classify a chant by its mode, two things must be known. The first is how the tonic is related to the diatonic genus as a whole.

The second is how the dominant is related to the tonic.

Our octave, divided in the diatonic genus, begins with a fourth which is itself divided into a tone, a semitone, and another tone. The tonic is in one of these four places. In modes I and II it is in the first place (the whole string), in modes III and IV it is in the second place (a tone above the whole), in modes V and VI it is in the third place (a minor third above the whole), and in modes VII and VIII it is in the fourth place (a fourth above the whole).

The word "third" is ambiguous. A "major third" is two tones. A "minor third" is a tone and a semitone. Sometimes the context makes it clear which one we mean, in which case we can simply say "third."

The next principle is the relation of the dominant to the tonic. In the odd modes, the dominant is a fifth above the tonic. In the even modes, the dominant is a third below the dominant of the corresponding odd mode.

A semitone is not a stable interval. The lower note tends naturally towards the higher one. This tendency leads to the following modification of the rule to classify modes. If the interval which determines the dominant is immediately below a semitone, then the pitch above the semitone becomes the dominant. This exception applies in Modes III, IV, and VIII.

The odd-numbered chants usually range through the octave above the tonic. The even-numbered chants, on the other hand, range through the fifth above and the fourth below the tonic. We will see this in examples below.

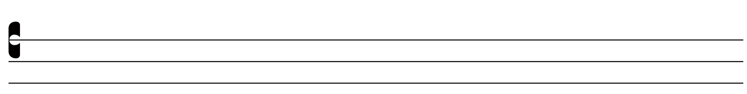
This is why we left room below the "whole" when we divided our monochord.

21.3 Notation

The notes of Gregorian chant are written on four horizontal lines. Both the lines and the spaces between the lines are used to mark notes. To move from one line to an adjacent space, or vice versa, is to move by either a tone or a semitone. Higher pitches are higher up, and lower pitches are lower down.

There is a small marking, known as a clef, at the beginning of each line of music. It indicates how the lines and spaces relate to the diatonic genus. There are two kinds of clef. Each of them indicates the place of a note immediately above a semitone. The whole octave, in the diatonic genus, has two semitones. The two clefs mark the upper note of each of these.

The clef that is used more frequently looks like this.



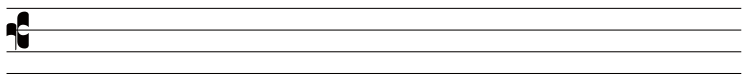
It is called a do clef. It marks the top of the upper semitone in the

The "do" is pronounced as in female deer, not act or make.

octave. That means that the whole, the note that begins our octave in the diatonic genus, is the bottom line of the staff in this case.

The do clef is not always on the top line.

Sometimes a clef like this is used.



It is called a fa clef. It marks the top note of the lower semitone in the octave (i.e., a third above the whole). In this case the fa clef is indicating that our octave in the diatonic genus begins on the second line from the bottom.

21.4 *Examples*

The following examples exhibit characteristic features of their respective modes. You do not need to know musical notation in general and Gregorian chant notation in particular to understand the examples that follow. Some details, such dots, lines, and variations in the shapes of the notes, indicate rhythmical changes and lengthening of notes. Do not worry about those details. These principles are sufficient.

1. The clef determines the location of the semitones within the lines and spaces.
2. Notes proceed from left to right.
3. When two notes are directly above and below each other, the lower one comes first.
4. When there is a long stroke, that looks like a pen was not lifted from the page, this indicates a note at the starting point of the stroke followed by a note at the ending point of the stroke.
5. The syllable that is sung with certain notes is written under those notes. There are often elaborate passages where many notes correspond to a single syllable. The vowel of a syllable is continued through all notes until the next place that a syllable is written.

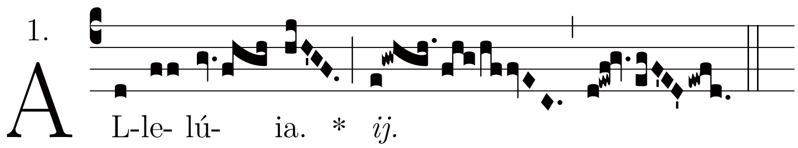
An exception to this rule will be noted in an exercise. It is important to realize that you are being given a good, but rough, outline of a complex body of music and notation.

If you can find someone who knows how to sing the music, it will be helpful. You can also find recordings easily.

21.4.1 *Mode I*

The first mode has the tonic placed at the very bottom of our octave, on the whole string. The dominant is a fifth above the tonic. Here is an example. This is the Alleluia for the Sixteenth Sunday after Pentecost.

The liturgical dates refer to the calendar used throughout the Roman Catholic Church prior to the 1960s.



The letters *ij* refer to a repetition that occurs in the chant. They are not part of the text that is sung.

The number 1 indicates the mode. Observe that the do clef is placed on the top line. The tonic is an octave beneath the note above this, or equivalently a seventh beneath the place of the do clef. This means that the tonic is on the bottom line. The dominant, which is a fifth above the tonic, is on the second line from the top.

The chant both begins and ends on the tonic. It ascends to the dominant in the third syllable of the word "Alleluia" and moves around this pitch, both above and below, before eventually returning to the tonic at the conclusion. There are four times that the melody is on the dominant, departs from it for one note, and then returns to it. See if you can find them all.

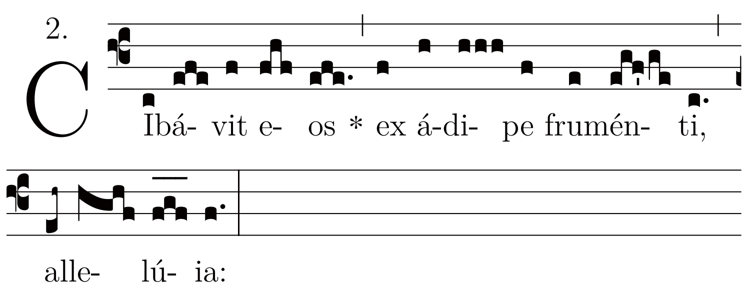
What has been given is in fact only a part of the whole chant, but it has the key features.

21.4.2 Mode II

The second mode has its tonic in the same place as the first mode, on the whole string, but the dominant is nearer, only a third above. Here is an example. It is the Introit for the Feast of Corpus Christi.

The chant begins this way.

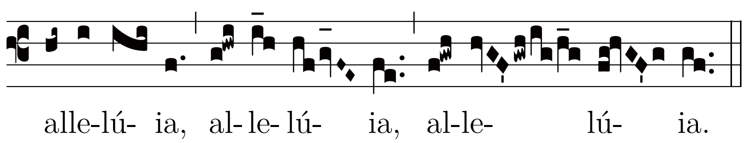
The Introit is the chant at the beginning of Mass as the priest enters and begins.



The clef is a fa clef rather than a do clef. The fa clef indicates the note above the semitone in the lower fourth of our octave, i.e., a minor third above the whole. Thus, the tonic is two notes below the clef, on the second line from the bottom. The chant begins well below this note, and also returns to that low pitch. Recall that in even modes the chant includes notes significantly below the tonic, and does not go much more than a fifth above it.

In the first syllable of the word *eos* we see the characteristic notes, the tonic and dominant, of the second mode. The sequence of notes for this syllable consists of the tonic, the dominant, and then the tonic again. On the second syllable of *adipe* the three notes in a row on the dominant indicate a long hold. This is a place where energy develops. The energy is not fully released in the word *frumenti*, as that word only briefly touches the tonic before descending below. Finally, in *alleluia*, the first two syllables reiterate the energy of the dominant before resolving in the tonic.

Some of the chant has been omitted. Here is the conclusion.

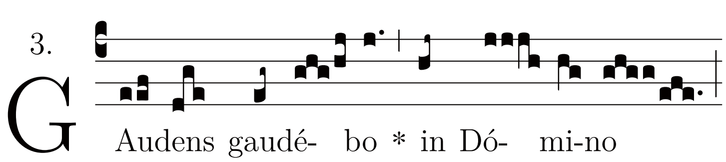


Here, unlike in the section earlier, we do pass above the dominant, but only by a semitone, to the point a fourth above the tonic. In the final *alleluia* we see movement between the tonic and dominant, movement around the dominant, and finally resolution in the tonic.

Remember that each clef marks the lower note of a semitone interval.

21.4.3 Mode III

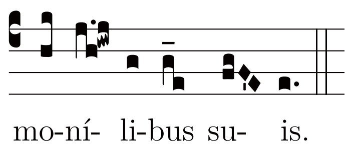
The third mode is the first mode in which the instability of the semitone affects our analysis. The tonic is immediately below the semitone of the lower fourth of our octave. A fifth above this is the note immediately below the semitone of the upper fourth of the octave. The proximity of that note to the one immediately above means that the dominant becomes the sixth above the tonic. Here is an example. It is the Introit for the Feast of the Immaculate Conception.



The do clef is on the top line. In the first mode, this indicated that the tonic would be on the bottom line. Now, in the third mode, the do clef on the top line indicates that the tonic is on the space immediately above the bottom line. The chant begins on this note.

After beginning on the tonic, the chant climbs up to the dominant at the end of the second word, the word *gaudebo*. It climbs to that note again in the second phrase, the phrase beginning with the word *in*, and develops a sense of expectation in the long hold on the first syllable of *Domino*. This resolves in the tonic at the end of the word, which concludes the chant's initial phrase.

The intermediate portion of the chant is omitted. If you find a copy of the whole chant you will see that some of the omitted phrases move around the dominant without resolving in the tonic. This gives those phrases a provisional, incomplete sense. Here is the conclusion.

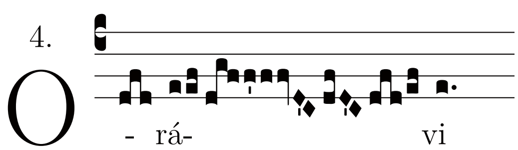


This final word begins near the dominant and ascends above it. It then lingers on it before descending to the tonic.

21.4.4 Mode IV

The fourth mode, like the third mode, has its tonic immediately below the semitone in the lower fourth of our octave. Since Mode III has a higher dominant, the dominant of Mode IV is a fourth above the tonic. Chants in this mode range above and below the tonic. Here is an example. It is the Offertory for the 17th Sunday after Pentecost. This is the first word.

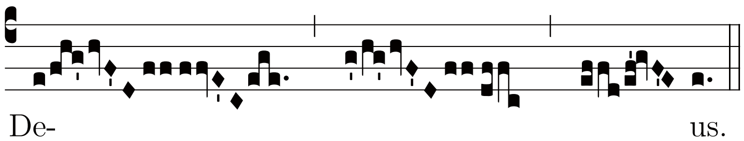
The pitch that is a third above the tonic can also be significant in chants of the fourth mode.



The do clef is on the top line, and the mode is the fourth, so the tonic is on the space above the bottom line. This is what we saw also in the third mode.

The chant does not begin on the tonic. Instead it begins below it. In this first word it only briefly reaches near the dominant and spends a good bit of time below the tonic, reaching a third below it at the space beneath the bottom line.

The chant ascends as high as the space above the top line. That is a seventh above the tonic. The chant also repeatedly revisits the third below the tonic, on the space below the bottom line. Here is the final word.



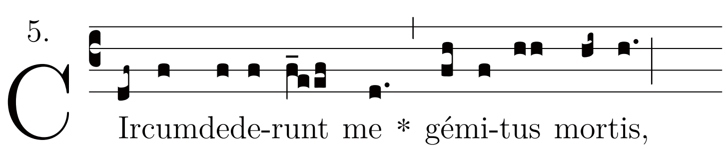
The small vertical lines (there are two of them here) intersecting the top line indicate phrases, musical and rhythmical units, within the larger composition. The first of the phrases in this final word rises from the tonic to the dominant, and descends below the tonic before concluding on it. This gives a sense of conclusion. The next phrase, the intermediate one, does not conclude in this way. It rises to the dominant and then descends well below the tonic while avoiding it entirely. This contributes suspense to the chant. That suspense is then resolved in the final phrase.

If you play the notes indicated above, or find someone to sing them to you, or hear a recording of the chant, you will see that what we call "resolution" in this case is somewhat subtle. It is less stable than the resolution in popular contemporary music. The note on which this chant concludes, its tonic, does indeed sound terminal, but it also lingers. The result is that it can bring about in the listener a kind of alertness. The thing is, in itself, concluded, but it has created in us a sense of expectation. The conclusiveness is restful but not lethargic.

21.4.5 Mode V

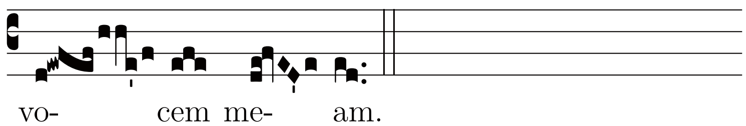
The fifth mode has its tonic immediately above the semitone in the lower fourth of our octave. As an odd mode, the dominant is a fifth above. Here is an example. It is the Introit for Septuagesima Sunday.

Septuagesima is the Sunday two and a half weeks before Ash Wednesday.



Observe that the do clef is now in a different place. We had seen it on the top line before. Now it is on the second line from the top. This means that the tonic for the fifth mode is on the bottom line. The dominant, then, is on the second line from the top, the line marked with the clef.

The intermediate part of the chant is omitted. Here is the conclusion.

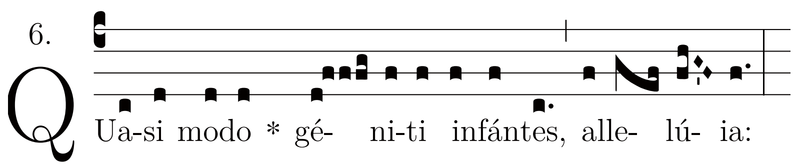


We see that, as always, the chant concludes on the tonic. Before this conclusion it reaches up once more, in the word *vocem*, to the dominant above.

21.4.6 *Mode VI*

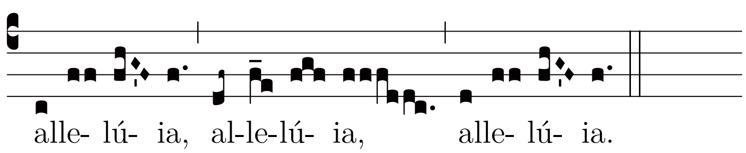
The sixth mode, like the fifth, has its tonic above the semitone of the lower fourth of our octave. Its dominant is a third above that. Here is an example. It is the Introit of Low Sunday, the Sunday immediately following Easter.

This day is now often called Divine Mercy Sunday as well.



The do clef is on the top line, so the tonic of the sixth mode is on the second line from the bottom. The dominant is on the second line from the top. In this early part of the chant we rest at and below the tonic, and only briefly touch the dominant in the middle syllable of *alleluia*.

Here is the conclusion of the chant.



In an even mode chant we expect to see substantial use of notes beneath the tonic, and that is the case here. The dominant is touched only briefly. This chant as a whole centers around the tonic. The concentration of the chant in a small range around the tonic gives it a sense of stability.

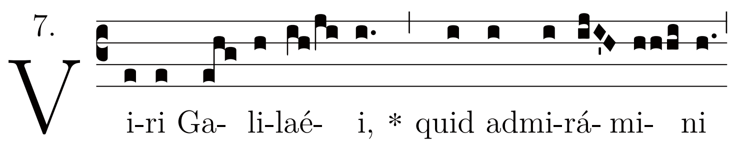
This chant, used at the beginning of Mass, begins with the words *Quasi modo*. The Sunday can also be called "Quasimodo Sunday." This is the source of the name of the character Quasimodo in Victor Hugo's novel *The Hunchback of Notre-Dame*, since he is found on that day.

21.4.7 *Mode VII*

The octave, as we divided it in the diatonic genus, consists of two fourths separated by a tone, and each of the fourths is divided into a tone, a semitone, and a tone. The four different places in the lower fourth of the octave led to four classes of chant, according to the possible places of the tonic. Now, in the seventh mode, we are at the final place for the tonic, at the fourth above the pitch taken to be whole.

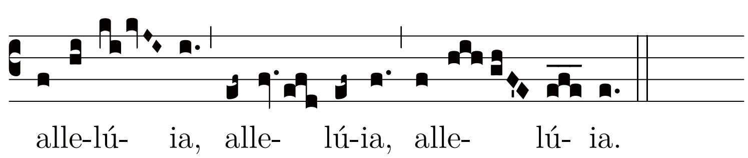
The division into four classes by the position of the tonic is further divided by the interval relating the tonic to the dominant.

Here is an example of the seventh mode. It is the Introit for the Ascension. These are the first words.



The do clef is on the second line from the top. This means that the tonic is on the space above the bottom line. This is where the chant begins. The dominant is a fifth above, on the space below the top line. This is where the word *Galilaei* ends.

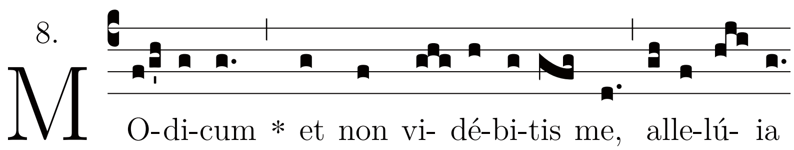
In the intermediate portion of the chant, the notes remain for the most part near the dominant, never rising above it, and then descend again to the tonic. Here is the conclusion, with the word *alleluia* repeated three times.



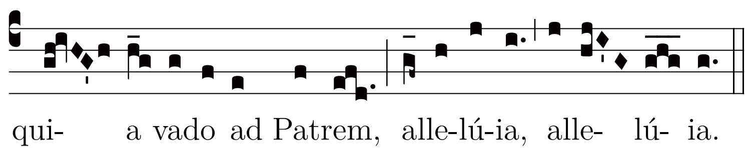
The first *alleluia* soars above the dominant by a third, to a point not yet reached in the chant. This dramatic *alleluia* is brought to a close at the dominant, giving it a partial sense of rest. The second *alleluia* is much lower, and concludes at a note just above the tonic. This produces an sense of incompleteness, of something needing to be resolved. That resolution is effected by the final *alleluia*, reaching up once more to the dominant before descending almost step by step to the tonic.

21.4.8 *Mode VIII*

The final mode, the eighth mode, is exceptional, like the third and fourth modes. The third below the dominant of the seventh mode is the lower note of a semitone interval, so the dominant of the eight mode is raised. Thus, in the eighth mode the interval between the tonic and dominant is a fourth. Our example of this mode is the Communion of the Third Sunday after Easter. Here is the beginning.



The do clef is on the top line, so that the tonic is on the middle space. The first word, *Modicum*, moves below and above this note before concluding on it. The dominant is a fourth above, at the top line, the line marked by the clef. Here is the conclusion.

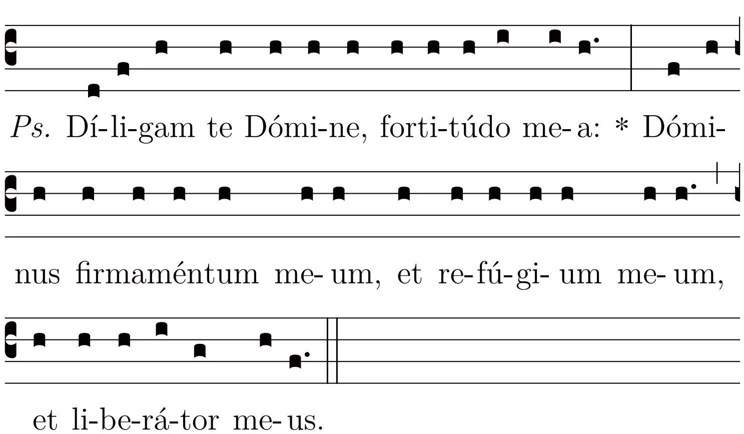


The phrase *vado ad Patrem* begins from the tonic and descends a fourth below. Even modes use more notes below the tonic, and we see that here. The word *alleluia* is repeated twice. The first *alleluia* touches the dominant but then concludes below it, lending a sense of expectancy. The second *alleluia* moves from the dominant to the tonic.

21.5 Psalms and Reciting Tones

Some chants of the form given above are complemented by psalms that are chanted in a simpler way. These psalms clarify the significance of the dominant, which is the main pitch on which the psalm is sung. The dominant can then be called the "reciting tone."

Here is an excerpt of the psalm that accompanies our Mode 5 example, the Introit *Circumdederunt*.



Recall that the tonic is on the bottom line and the dominant is on the second line from the top. The psalm is chanted almost exclusively on the dominant, the reciting tone. There are small bits of ornamentation at the end of phrases.

Observe that the psalm does not conclude at the tonic but instead at an intermediate point. After this psalm is sung, it is followed by the phrase *Gloria Patri . . .* , and then the antiphon *Circumdederunt* is sung again. Thus, the psalm does not need to conclude on the tonic. It is not the end of the whole chant.

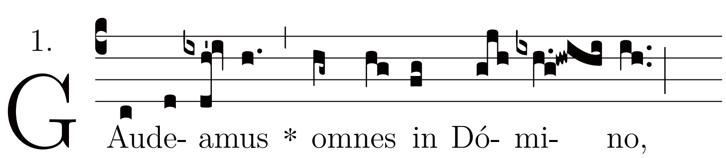
21.6 *Exceptions*

The preceding introduction to Gregorian chant omits some exceptional circumstances. One of them, the use of accidental notes, is quite common. A second one, transposition, occurs occasionally. A third one, a psalm tone does not fall neatly into our classification, is much rarer.

21.6.1 *Accidentals*

At times in Gregorian chant, the composer elects to vary an interval. The situation in which this occurs is when there would ordinarily be a tone separating two pitches, and instead the higher note is lowered to be only a semitone above the lower one. Such a note is called an "accidental."

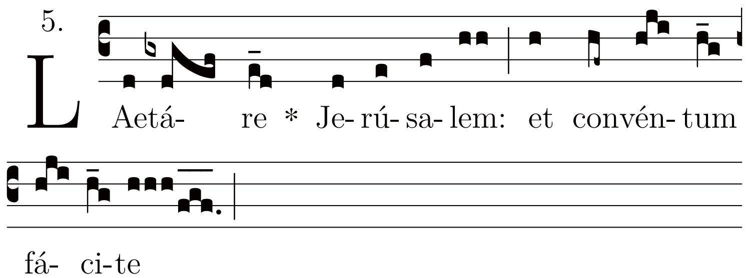
One example of accidentals is in a common pattern that begins many chants of the first mode. The Introit for All Saints Day is an example.



The symbol that looks like the letter "b" indicates that the note in that place has been lowered. The do clef is on the top line. This means, ordinarily, that there is a whole tone separating the second line from the top and the top space. Due to the accidental marking, the interval between those two notes is reduced to a semitone.

When the accidental (or "flat") symbol is used in Gregorian chant, it applies only to the word in which it occurs. If additional notes are to be modified, the symbol must be reused. That is why the symbol for an accidental occurs twice in this excerpt. The second occurrence is in a different word than the first one. It is not the case that the notes of the top space are always lowered in this chant. Later, in part of the chant not shown here, this note occurs with its natural value.

Here is a second example. It is the Introit for the Fourth Sunday of Lent, often known as Laetare Sunday.



In the opening word, *Laetare*, there is an accidental that lowers the note indicated by the middle space. That lowering is only applicable to the word *Laetare*. When the middle space is used to mark notes in the words *conventum* and *facite*, those notes take their ordinary, unmodified value, a whole tone above the note of the line below.

21.6.2 Transposition

There are chants in which the mode and clef do not align as clearly with each other. You will explore an example of this in an exercise. This is called transposition. In the case of a transposed chant, the tonic and dominant are related by the interval determined by the mode, but they are placed at a different point in the octave.

21.6.3 Wandering Tone

In the elaborate antiphons seen earlier, the tonic and dominant play clear roles, but many other notes also enter the chant. In the simpler psalm tone, by contrast, we see that the dominant is repeated, almost exclusively, in its use as the reciting tone.

There is a pattern for psalms that has two reciting tones, and thus does not match up neatly with what we have said above. Here is the example, from Vespers of Sunday.

Vespers is Evening Prayer.

The note on the second line from the top serves as the reciting tone for the first part of the phrase. The note on the middle space, on the other hand, is the reciting tone for the second part of the phrase.

This tone is called *peregrinus*. It is a "wandering" tone, since it changes its reciting pitch. The wandering tone is not common. It is included here to indicate the variety present in Gregorian chant, reminding us that our account is not comprehensive.

21.7 Solfege

You might be familiar with these terms: *do, re, mi, fa, so, la, ti*. This way of naming notes within an octave, called solfege, comes from Gregorian chant. The chant below is the first verse of a Hymn from the Nativity of St. John the Baptist.

The date of this feast is June 24.

This is in mode 2, and the fa clef is on the second line from the top. Observe that the first word of each phrase occurs one note above the previous one. These are *ut*, *resonare*, *mira*, *famuli*, *solve*, and *labii*. If you take the first syllable of each of these words you obtain, roughly, the familiar "do, re, mi, . . . " The two exceptions are that "ut" has become "do," and the note for "ti," the space above the top line, does not occur in the chant. The "ti" was added so that each note in the octave would have a name.

The phrases are divided by the vertical bars of various lengths.

The term "major scale" refers to the sequence of notes beginning with "do." It also leads to a division of the octaves in tones and semitones. This division is related to ours by shifting everything by one tone. The reason to adopt the division of the octave that we used, instead of the major scale, is that it relates more immediately to the classification of chant by modes. We also used a symmetrical division of fourths, which is pleasantly uniform. The major scale requires an asymmetrical division of fourths above the whole and below the octave.

Observe that the names “do clef” and “fa clef” now make sense. They refer to notes named by solfege. The do clef is in the place of the note do, which is at the word *Ut* and also one octave above it. The fa clef is at the place of the note fa, as we see in the chant above at the beginning of the word *famuli*.

The word *famuli* is from *famulus*, which means servant or slave.

21.8 Exercises

Exercise 1

Write a staff with four horizontal lines, and draw a do clef on the top line. Indicate the place of the two semitones on the staff.

Exercise 2

Write a staff with four horizontal lines, and draw a do clef on the second line from the top. Indicate the place of the two semitones on the staff.

Exercise 3

Write a staff with four horizontal lines, and draw a do clef on the second line from the bottom. Indicate the place of two semitones on the staff. Put one below the do clef and one above the do clef.

Exercise 4

Write a staff with four horizontal lines, and draw a fa clef on the second line from the top. Indicate the place of the two semitones on the staff. Both are below the fa clef.

Exercise 5

Write a staff with four horizontal lines, and draw a fa clef on the top line. Indicate the place of the two semitones on the staff. Both are below the fa clef.

Exercise 6

Play each piece from the chapter on the monochord. Remember that, when notes are stacked, the lower one is played first. Do not worry about dots and horizontal lines. They refer to lengthening. You can get a sense of how this works by listening to a recording.

Exercise 7

Mark the flatted "ti" on your monochord. Do this by ear, or find a way to do it mathematically.

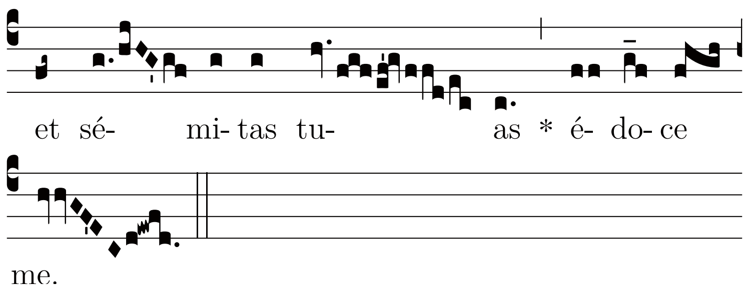
Exercise 8

Explain why it is useful to have a fa clef. Where would a do clef need to go if it were expressing the same position of the notes?

Exercise 9

Here is the conclusion of a chant. The chant is the Gradual for the First Sunday of Advent. It begins with the word *Universi*.

The Gradual is chanted after the Epistle and before the Alleluia. In all the remarks about the Mass, we refer to the form that was in widespread use before the 1960s.



a.) Identify the tonic.

Where does the chant end?

b.) Determine the range, relative to the tonic, of notes used in this portion of the chant.

c.) Observe the structure of notes away from the tonic.

d.) Use parts b. and c. to identify the dominant.

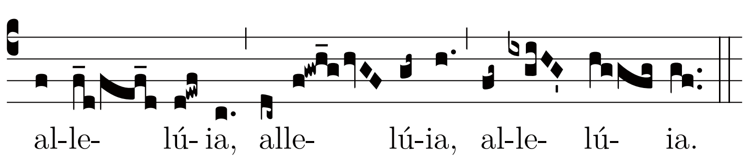
e.) Use the tonic and dominant to identify the mode.

f.) Does the word *tuas* conclude in a way that is restful, or anticipatory? Use your knowledge of the tonic and dominant and their roles to answer the question.

g.) Play this portion of the chant on the monochord and check your answer to the previous question.

Exercise 10

Here is the conclusion of the Communion for the Mass of Easter Day. This chant begins with *Pascha nostrum*.



a.) Identify the tonic.

b.) Determine the range, relative to the tonic, of notes used in this portion of the chant.

c.) Observe the structure of notes away from the tonic.

d.) Use parts b. and c. to identify the dominant.

e.) Use the tonic and dominant to identify the mode.

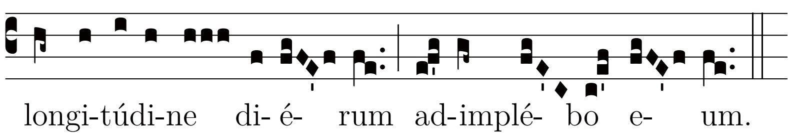
f.) The first *alleluia* and the second one are each melodically incomplete, but they are so in different ways. Explain the ways that they build a sense of anticipation prior to the concluding *alleluia*.

g.) Play this portion of the chant on the monochord and check your answer to the previous question. Note that the first notes of the second *alleluia* are special. The notes for the syllable *al* should be played with the top one first, and then the bottom one.

This is one of the exceptional instances mentioned earlier, in which the top note comes first rather than the bottom.

Exercise 11

Here is the conclusion of a chant. The chant is the Introit for the First Sunday of Lent, and begins with *Invocabit*.



a.) Identify the tonic.

b.) Determine the range, relative to the tonic, of notes used in this portion of the chant.

c.) Observe the structure of notes away from the tonic.

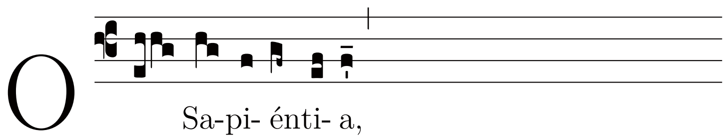
d.) Use parts b. and c. to identify the dominant.

e.) Use the tonic and dominant to identify the mode.

f.) Use the terms "tonic" and "dominant" to explain how anticipation is built and then resolved in the words *longitudine* and *dierum* in the penultimate phrase.

Exercise 12

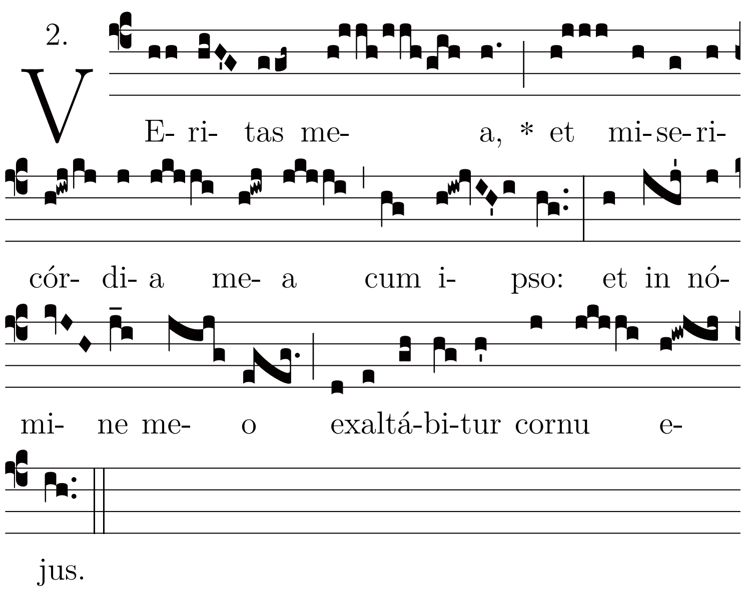
At the end of Advent, just before Christmas, there are a series of chants used in the evenings. Each of them begins with "O" as the Savior is addressed by various names. These chants are the basis for the hymn "O Come, O Come, Emmanuel." Here is the begining of the first of these chants.



Phrases can end in a variety of places, but this one in fact concludes on the tonic. Do your best to determine the mode of the chant from this small initial piece alone. You will see the conclusion of the chant in a later exercise, but you should not look ahead to that part until you have done your best with this initial part alone.

Exercise 13

Here is a chant with the fa clef in an unusual place. This is the Offertory for the Mass of a Confessor who is not a Bishop.



a.) Mark the places that semitones occur. Both are below the marked clef.

b.) Determine the largest interval above the tonic used in the chant.

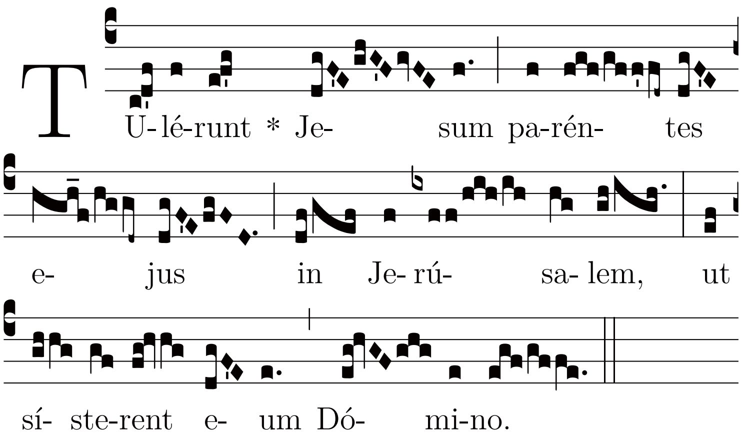
c.) Determine the largest interval below the tonic used in the chant.

d.) Discuss the melody of the word *ipso*, which occurs immediately before the full vertical bar. What melodic device lends motion to this intermediate conclusion, so that we are led to the beginning of the next phrase?

Note that while the notes are close together, the higher one of the final syllable is to the left of the lower one, and thus comes first.

Exercise 14

Here is the Offertory for the Feast of the Holy Family.



a.) Identify the tonic.

b.) Determine the range, relative to the tonic, of notes used in the chant.

c.) Observe the structure of notes away from the tonic.

d.) Use parts b. and c. to identify the dominant.

e.) Use the tonic and dominant to identify the mode.

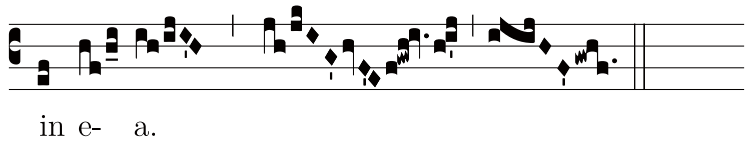
f.) Identify the ways that *ejus* and *Jerusalem* conclude relative to the tonic and dominant, and discuss the musical effect.

g.) Observe that notes other than the dominant and tonic occur with greater frequency than they do. Consider, for example, the word *parentes*.

Exercise 15

Here is the beginning of *Haec dies*, the Gradual of Easter Day.

Observe that the chant is in the second mode, but is placed relatively high in the octave, near the do clef. Here is the conclusion.



The mode was marked as 2, yet the chant is concluding near the do clef. This is unusual, and is a case of transposition.

a.) Determine the location of the semitone.

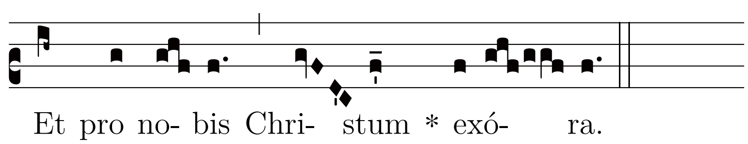
b.) Determine the interval between the last two notes of the chant. These are the dominant and the tonic.

c.) Compare the results of parts a. and b. to the usual form in which mode 2 is presented. You should conclude that in both cases the tonic and dominant are separated by a minor third (the interval composed of a whole tone and a semitone) and that the whole tone is immediately above the tonic.

d.) Check modes 4, 6, and 8. What is the interval from the tonic to the dominant? If it is a minor third, where is the whole tone? Discuss in connection with this example.

Exercise 16

Here is the end of the Solemn Tone of the *Ave Regina Caelorum*. This chant is listed as being in Mode VI.



This is another example of transposition. Ordinarily the do clef does not mark the location of a tonic. The word *exora* includes the dominant as its highest note and the tonic as its lowest (and final) note.

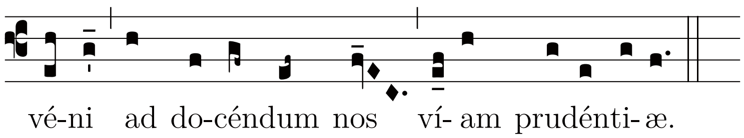
a.) Look back at an example of a chant in Mode VI and determine the interval between the tonic and dominant in that mode.

b.) Identify the interval between the tonic and dominant in the *Ave Regina Caelorum*, given here.

c.) Explain why it is reasonable to call this a Mode VI chant despite the unusual location of the tonic relative to the clef, using the previous two parts. Compare with the previous exercise.

Exercise 17

Here is the conclusion of the antiphon *O Sapientia* that you examined earlier.



Use this portion of the chant to confirm or correct your earlier determination of the mode.

Part IV

Astronomy

Dilexi decorem domus tuae.

22 Observation

Look at the sky, day and night. You must do this regularly before you attempt to learn about mathematical accounts of astronomical motion. Complete the exercises of this chapter while you study the first three parts of the book. Only once you have completed the observational exercises, and the first three parts of the book, should you continue beyond this chapter.

This book will assume the observer is in the Northern Hemisphere at a somewhat northerly latitude. It can still be used, with some minor adjustments, if you live in the Southern Hemisphere or near the equator.

22.1 Exercises

All of these observational exercises should be completed before you continue to the rest of this book. Read through all of the instructions. Some of the exercises can be accomplished at the same time. You do not need to do them in a specific order. Read the instructions carefully and make a plan, using a calendar, for how you will complete all of them over the course of a few months.

Exercise 1

This exercise will take between two weeks and a month, and involves observations made at midday. It is best to do it in the spring or the fall. You will record the shadow length of a fixed object.

What is the purpose of this exercise? The daily path of the sun through the sky varies throughout the year. Sometimes it is higher in the sky, and sometimes lower. You will pay attention to the way the sun's path changes and make some measurements to see how quickly it changes.

Place a stick in the ground, or find a pole making a clearly delineated shadow whose end you can mark. Use paint or chalk to mark the end of the shadow at noon. Measure the length of the shadow with a measuring tape. Do this again after a couple of weeks. If possible, continue doing this every couple weeks. It is tricky if you do this in June or December, because at that time of year it is harder to

The stick or pole you use should be something like a yard or meter tall, or taller.

see the changes. Try to do it near March or September if you can. Make a clear record of your observations.

Exercise 2

This exercise requires regular observations over the course of a month, every couple of days. Some are during the day and some are at night.

What is the purpose of this exercise? The moon varies in its appearance each month. Sometimes it is a thin crescent, sometimes it is a quarter, sometimes it is full. You will pay attention to what the moon looks like and where it appears in the sky in relation to the sun. This exercise will help you to know and remember which lunar phase is a first quarter moon and which is a third quarter moon.

Over the course of one month, observe the appearance of the moon and the relationship between the moon and the sun. Every two or three days, look for the moon, observe its shape, and record where you saw it in the sky, and at what time.

Sometimes the moon will be visible during the day. Other times it will be visible at night. You need to keep your observations spaced fairly closely together, within the span of a couple of days, so that you do not lose track of it. The place you can see it after a couple of days will be fairly close to where you saw it last.

Exercise 3

This exercise takes place over the course of a single night. It is easier to do in winter when nights are longer. It is good if you can make an observation in the middle of the night, but you should at least make the first observation and the final one, given in the first and third steps.

What is the purpose of this exercise? When we consider mathematical models for motions of the sun, moon, and stars, we will think of them as moving within a huge sphere around the earth. By looking at how the stars move throughout a single night, you will get a sense of why this spherical model is a good one.

1. Go out shortly after sunset, when stars are first visible, and look to the north, so that the sunset is on your left. Record the positions of some of the brighter stars that are visible.

In the southern hemisphere, you should look to the south, so that the sunset is on your right.

2. Wake up in the middle of the night and go outside again to look at the stars to the north. Make a record of their positions.

3. Wake up before dawn and look once more to the north. Make a final record of their position.

4. Write a brief explanation of what you see in your observations.

Exercise 4

This exercise takes place over the course of about a month. You will make a series of observations of the stars near sunset. You should make the first and last observations at the same time of day. The intermediate observations should be around that time, but need not be exact. The purpose of the intermediate observations is simply to help you keep track of the things you saw at the beginning so that you can compare them to what you see at the end.

It is fine if you simply use an ordinary clock to note the time of your observations. If you want to be more exact, though, you can use time relative to sunset. This variation is described below, after the main instructions.

What is the purpose of this exercise? The nightly motion of the stars is, to our eyes, quite uniform. So is the daily motion of the sun. These motions are not quite the same, however. This exercise helps you to see the way that they are slightly different, and to describe one using the other.

If you begin this exercise in the spring, wait at least an hour after sunset. You need to leave some room, since days will be getting longer.

When you look to the south in the evening, the sunset is to your right. When you look to the north, the sunset is on your left.

1. Go out somewhat after sunset, once stars are visible. Make a sketch of the stars that are visible when you look to the south. Make another sketch of the stars visible when you look to the north. Only look at a few bright ones, but mark their positions carefully. If you know names of some constellations it will be helpful, but you do not need to know them. Include in your sketch some nearby landmarks on the horizon, and mark the time at which you made the observation.

When you make the intermediate observations, look back at your notes from the first observation. You must keep track of the stars and patterns that you originally thought were notable.

2. Repeat the observations about one week later, around the same time, though you do not need to be exact.

3. Repeat the observations about two weeks after the initial observation, again around the same time, though you do not need to be exact.

4. Repeat the observations about three weeks after the initial observation, again around the same time, though you do not need to be exact.

5. Repeat the observations about four weeks after the initial observation. Here you should be more exacting. Do this at the same time as your first observation.

6. Write a description of how the apparent positions of the stars and the sun have changed in relation to each other. Think about what it would be like if you had taken a panoramic photograph in each case. Each time, the sun is just over the horizon. How does

the first (imaginary) photo compare with the second one? Are the stars in the same place? Farther east, farther west? Something else?

Variation: If you wish to make this series of observations more exactly, do not use the time given by a clock. Instead, find the time of sunset for your location. Make your observations at the same number of minutes after sunset (e.g. 60 or 90 minutes after sunset). Over the course of a month the time of sunset could change substantially, so you can account for this by observing relative to the sunset rather than the clock. Here is an example. Suppose that the first observation was made at 9:15 pm, on a day when the reported sunset was at 8:10 pm. That means the observation was 65 minutes after sunset. Suppose further that one month later, when making the final observation, the reported time of sunset is 7:50 pm. Then the observation should be made at 8:55 pm, since that is 65 minutes after sunset.

It is easy to find sunset times online.

Exercise 5

This exercise has you watch the relation of the sun and moon over time. The way that you complete the exercise depends on the time of year that you do it. If you have the chance to complete both parts, do so. It is best to do this exercise in one of the eight months listed below.

What is the purpose of this exercise? Both the sun and moon vary in their height above the horizon. The way that they vary is related, but the two motions are distinct. This exercise guides you to see how they are related.

It is best to do the exercise in one of the months given below. If for some reason you are unable to do this, do the summer/winter observation (the first one listed).

If you complete the observation in June, July, December, or January:

1. Find out the date when the full moon will appear. The moon does not need to be exactly full, but should be within one or two days of it.

2. Go out at midnight, or near midnight, and observe the place of the moon in the sky. More specifically, record your sense of its height above the horizon, around midnight.

3. Make a record of what you see and the date of your observation. Compare the place of the moon in the sky **at midnight** to the place of the sun **at noon** on the day before or the day after.

If you complete the observation in March, April, September, or October:

It is fine to observe the third quarter moon first and then to observe the first quarter moon second. The order does not matter. You can find the dates of the moon's phases online. The terms "first" and "third" are common and you will find them, even if you do not yet understand what they mean.

1. Find out when the moon will next be at its first quarter and its third quarter.
2. When the moon is at its first quarter go outside at **sunset** and observe its place in the sky. How high is it above the horizon?
3. When the moon is at its third quarter go outside at **sunrise** and observe its place in the sky. How high is it above the horizon?
4. Make a record of your two observations.

Exercise 6

This is an optional exercise. If you are making a long trip while studying this book, and your trip will take you to the north or south by a significant amount, you can try this exercise.

What is the purpose of this exercise? It will help you to perceive the sphericity of the earth through simple measurements, and understand terrestrial latitude.

If you wish to make the observation in an especially precise way, look up what is called "solar noon" for each location and make the shadow measurement at solar noon each time.

1. Shortly before traveling, within a few days of your trip, record the length of the shadow of an object at noon. Record the height of the object as well.
2. During your trip, make the same observation, recording the length of the shadow of an object at noon. The object does not need to be the same as the original one, but you must record its height as well.
3. When you return from your trip, repeat the measurements. Depending on the duration of the trip, you might get substantially different results before and after travel. Be sure to note the dates of all measurements, at home and while away.

22.2 *Principles*

Do not read this section, or any further in the book, until you have completed the preceding exercises.

You have watched the sky, seeing the sun and moon over the course of months. The following principles should make sense to you when you think about what you have seen.

Principles of Solar and Lunar Motion

1. At noon in the summer the sun is high in the sky, and at noon in the winter the sun is low in the sky.

2. The moon falls back relative to the sun, considering their daily motion from east to west.

3. The daily paths of the sun, moon, and stars through the sky are like arcs of circles.

4. The sun falls back relative to the stars, considering their daily motion from east to west.

5. The sun and moon travel in roughly the same places among the stars.

The first three are especially important. The last two might be more difficult to grasp if you live in an area where it is hard to see the stars, but you should still have some sense of them from your observations.

We will now discuss how those principles follow from what you have seen. The principles have been known for many years, by people who saw the same things that you did.

22.3 *Solar Declination*

The term "declination" is related to the height of an object above the horizon. The precise meaning will be clarified soon. The rough sense, in the case of the sun, is that it refers to how high the sun is above the horizon at noon.

You will see that "declination" is independent of the location of the observer, while the sketch given here depends on where the observer is on earth.

Over the whole year the sun is sometimes high in the sky at noon and sometimes low in the sky. The change is gradual. In the Northern Hemisphere, the sun is highest at midday in late June, and lowest at midday in late December.

This statement is not quite accurate for those who are in the Northern Hemisphere but very close to the equator. Details will come later.

If you made observations of shadows in the spring, the earlier noon shadows were longer, and the later noon shadows were shorter. This is because the sun was higher in the sky at noon on the later days, the days that were closer to June. If you made observations of shadows in the fall, on the other hand, the earlier noon shadows were shorter, and the later noon shadows were longer. This is because the sun was lower in the sky at noon on the later days, the days that were approaching December.

If you tried making the observations near late June or late December, you might have had trouble distinguishing the length of the shadows from one measurement to the next. This is for two reasons. One is that the sun's height at noon changes much less from day to day at those times. Another is that it could be that the shadow length did not change. For example, if you made a measurement on June 14, and then again on June 28, the shadows were probably the same length. This does not mean that the sun's height above the horizon was not changing. Instead, from June 14 to June 21, the sun got a bit higher each day, and then from June 21 to June 28 the sun got lower each day, returning to roughly the same spot as it had been previously. The small intermediate changes are harder to measure than the larger changes in spring and fall. You can measure them if you are careful, though.

The day June 21 is not exact.

22.4 Lunar Elongation

The term "elongation" is used here to refer to the extent to which the moon's rising and setting times differ from the sun's on a given day. It has a slightly more precise meaning that is not important right now.

What you observed when watching the moon during a month is that the moon falls back relative to the sun, or alternatively the sun gains on the moon. What does this mean more concretely? Suppose that on one day the moon rises around noon. This is the first quarter moon. On the next day, at noon, the moon will not yet have risen. It will not rise until later. The amount of time varies, but it will be at least about half an hour and could be more than an hour later, and in any case will always be later. This is what it means to say that the moon falls back relative to the sun. In one day, 24 hours, the sun has made one trip through the sky, but the moon has done slightly less. Over time, these lesser motions add up, so that the moon proceeds through all its phases.

Daylight Savings Time can change the times of these things by an hour, as can your position within a given time zone. More precisely, "noon" here should be thought of as solar noon, the time when the sun reaches its highest point in the sky, for observers far from the equator.

In most of the Northern Hemisphere, the sun and the moon tend to be in the south. That means that if you look at them when they are at their highest point above the horizon, the east is to your left and the west is to your right. Thought of in this way, the moon tends towards the left, day by day, relative to a fixed place for the sun. That means, for example, that if the moon is near its highest point at sunset, which is a first quarter moon, then a few days later the moon will be more than a quarter. It will be a what is called a "waxing gibbous" moon. And a few days after that the moon will be even further "to the left" (in comparison to the setting sun), at which point it will be near full.

22.5 Circularity and Sphericity

In order to give a mathematical account of the various things we see in the sky, we need a simple geometrical frame in which to set the various heavenly motions all together. The fundamental framework that we will use is that of a sphere. What you have seen is that the positions of the stars in the sky change over time, throughout the night, but they remain the same in their relation to one another. They might be higher or lower relative to the horizon, but they are at the same distances from each other no matter how high or low they are.

These observations lead us to use a simple model: a sphere whose center is the earth. This sphere rotates each day, and the points about which it rotates are the celestial poles. There are two poles. You can look up and see one of them, and the other remains hidden unless you go to the earth's other hemisphere.

When a plane passes through the center of a sphere, it intersects the sphere in a circle, which is called a "great circle." There are many planes that pass through the two celestial poles. Each of these planes leads to a circle on the sphere, and every object in the sky is on such a circle. The angle from the pole to the object gives what is called the "declination" of the object. The things whose declination is $90°$, which are at a right angle from the pole, are on what is called the celestial equator. Stars on the equator rise exactly in the east and set exactly in the west. They are in the sky above for exactly twelve hours. The height they reach in the sky depends on the observer's position on earth.

There are other circles than great circles on a sphere. They come from planes cutting the sphere but not passing through the sphere's center.

We will begin to talk about angles using measurement in degrees. This will be covered in the next chapter.

A useful term arises from the notion of the celestial sphere. As you look up in the sky, you see the stars going around a pole. Consider that point, the pole, along with the point which is directly overhead (called the "zenith"). There is a great circle on the celestial sphere that passes through those two points, the pole and the one directly overhead. This circle gets a special name.

Definition 101. *The meridian is the great circle determined by the celestial pole and the point directly overhead.*

As we think about a plane through the center of the earth that intersects the celestial sphere in the celestial pole and the zenith, we can also think about where this plane intersects the surface of the earth. This intersection is a line that runs exactly north and south; it is a line of terrestrial longitude.

22.6 *The Sun and the Stars*

Different stars are visible at different times of year. The reason for this is that the sun is moving within the sphere of the fixed stars. The stars that are near the sun are not easily seen, since they are dim in comparison with its brightness. Those stars that are far from the sun within the celestial sphere, however, are visible. We see them at night.

The motion of the sun and stars can be described in two stages. The first stage is the simpler one. This first stage is the motion of the sun and stars from east to west through the sky. It is a uniform motion completing a whole circuit each day. You see the sun rise in the east and set in the west. You see stars rise in the east and move towards the west.

The second stage of the description is more complicated. The sun and stars to not move through the sky in the exact same manner. If they did, we would always have the same night sky. We do not, however. Different stars are visible at different times of year. In the course of a year, the sun returns to its same place among the stars. It is necessary to observe which way it passes through the stars in a given year. Does it overtake them? Do they overtake it? Your observations should make clear to you that the stars move with a slightly more rapid motion. Said in a different way, the sun tends to fall back relative to the stars. You saw that, with the passage of time, the stars visible after sunset moved further to the west. This is because over each 24-hour period, the stars revolve slightly more than the sun does. As the incrementally greater motion accumulates over time, it becomes visible in the advance of stars relative to the sun.

We will later discuss the definition of year with greater precision.

22.7 *The Ecliptic*

When we watch the sun, moon, and stars from one location, we observe some distinctions in their movements. A star always follows the same path through the sky, whether we watch it in the middle of the night, near sunrise, or near sunset. The sun, on the other hand, takes different paths throughout the year. In the summer it reaches higher points in the sky, and in the winter it remains at lower points, further from overhead. The moon's behavior is the most complicated of all. In a single month it sometimes takes paths high in the sky, and at other times it remains lower.

It turns out that the moon's behavior can be described fairly easily in terms of the sun's. For this, we need a definition.

Definition 102. *The ecliptic is the path in the celestial sphere made by the sun in the course of one year.*

The ecliptic is a great circle. There are four special points on this circle. Two of them are where the ecliptic intersects the celestial equator. The other two are the points that are furthest from the celestial equator. These points have special names and correspond to the four seasons, as we will see later.

The ecliptic lies at an angle to the celestial equator. People have observed over time that this angle is relatively consistent, and is slightly less than 24°.

By using the term "ecliptic" we can account for the varied behavior of the moon. The moon always travels in the ecliptic, roughly

We will later refine the statement that the moon travels in the ecliptic.

speaking. Consider a full moon in winter. The moon is directly opposite the sun, because it is full. The moon is in the place that the sun is in summer, since the sun is in opposite places along the ecliptic in winter and summer. Thus, you can see that a full moon in winter will be high in the sky, and a full moon in summer will, by the same kind of reasoning, be low in the sky. This sort of reasoning will be explained in greater detail in the chapter about the moon.

22.8 Terrestrial Latitude and the Earth's Sphericity

Whether or not you were able to do the optional exercise, it is important to learn how we determine the shape and size of the earth. If you did the exercise, you should have found that traveling to the north meant that the sun stayed lower in the sky, and traveling to the south meant that the sun rose higher in the sky. These observations, repeated by many people in different places and times, convince us that the earth is shaped like a sphere. The variation in the noontime height of the sun reflects the different angles that the surface of the earth has at those locations, in relation to the light emanating from the sun.

Consider a day in the spring or fall, when the sun is at neither the extreme of winter nor of summer. On such a day, an observer can see at noon how close the sun is to being overhead. This distance, measured as an angle, is called the "latitude" of the location. People who are at the equator see the sun directly overhead at noon, in spring and fall. Their latitude is $0°$. People who live somewhat far from the equator, but not at a polar extreme, see the sun not overhead, but towards the south (if in the northern hemisphere). The amount of displacement from overhead is their latitude. Someone at the north or south pole would see the sun exactly on the horizon.

22.9 *Additional Exercises*

Exercise 7

Copy this statement.

> The movements of the earth, sun, moon, and stars are complicated. In order to talk about them mathematically, we must simplify some things. One way to simplify is to think of a huge sphere, whose center is the earth, and which contains the sun, moon, and stars.
>
> This is not the only mathematical model of these motions. When people choose a mathematical model, they do so with the goal of making certain kinds of explanations and descriptions. There is an element of freedom when we make a mathematical model.

The spherical model is too simple to account even for the whole of this book. Later, when we talk about solar models and lunar parallax, you will see why.

After you copy it, read it aloud.

Exercise 8

Compare the five astronomical principles of Section 22.2 to the five postulates for geometry in Section 3.2. In what ways are they similar? In what ways are they different? Can you imagine adding additional principles beyond these? Could anything lead you to reject these principles, or refine them?

23
Plane and Spherical Trigonometry

Where in the sky is the sun? Where is the moon? Where are the stars? When we answer these questions we use angles, like when we speak of the elevation of the sun above the horizon at noon. In order to make careful mathematical accounts, we need a system of precise computation involving angles and spheres.

This is a challenging chapter with significant technical depth. You are ready to study it, but you must do so with confidence and energy. It will be hard, at the outset, to see why we take so much time to talk about angles and their chords. For now, simply trust that it is a good thing to do. We are assembling a set of tools, and once we have our tools ready, we will put them to substantial use.

This chapter establishes computational tools that make the motion of the sun and moon intelligible through the geometry you learned earlier. There are two important tasks. One task is to compute certain kinds of quantities. The second task is to explain relationships among certain kinds of ratios. Once both of these are done, we can put them together. If we know that a relationship holds between certain ratios, and we know all but one of the terms in the ratios, we can find the desired, missing term. We compute simple things directly, and we compute complex things by using our knowledge of simple things along with our knowledge of relationships among ratios.

You have probably done something like this in an algebra class.

The specific quantities that we will compute are lengths of chords. A chord is a line segment in a circle. The name "trigonometry" is given to the study of lengths of such lines.

The theorem that shows relationships among ratios is named for the mathematician Menelaus. We will investigate his theorem first with ordinary triangles. After that we will see how it is useful when thinking about arcs on a sphere.

23.1 Chords

If we have two points on a circle, these two points determine a couple of things. One thing is the section of the circle that lies between them. This is called an "arc" of the circle. Another thing is the line segment determined by those two points. This line segment is called a "chord."

Here is a basic question to ask. Suppose we are given an angle at the center of a circle. What is the size of the chord determined by this specific angle? The size of the chord is called the "chord" of the angle.

It turns out to be convenient to establish a measuring system when talking about chords. Rather than using arbitrary circles, we use one in which the diameter is understood to consist of 120 units.

Our first chord computation will be simple. What is the chord of $90°$? In Figure 23.1, with a circle centered at A, we are asking for the chord associated to the angle BAC. That chord is the length of the segment BC. Our convention is that the length of radius AB is 60 (i.e., the diameter is 120).

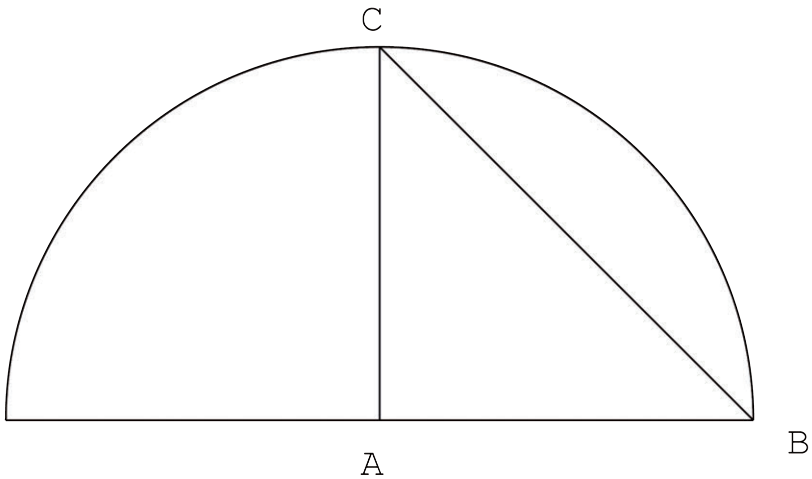


Figure 23.1: Chord of a right angle

Proposition 103. *The chord of* $90°$ *is approximately 85.*

Proof. Consider the right isosceles triangle ABC, with right angle at A. By the Pythagorean theorem the square on AB and the square on AC combine to equal the square on BC. The squares on AB and on AC are each 3600 square units, so the square on BC is 7200 square units.

The square of 84 is 7056, and the square of 85 is 7225, so the length of BC is between 84 and 85 units. It turns out to be closer to 85. □

Some of the details about which estimate is better are omitted. You can check the squares of 84 and 85 on your own, though.

23.1.1 *Finding Roots*

It might be unclear how we arrived at the numbers 84 and 85 in the preceding proof. Sometimes you can simply guess to find a root, but it is better to have a method.

Here is one method for finding square roots. In this section we will not worry about the difference between numbers and ratios, and instead will rely on the fact that you are familiar with fractions and their arithmetic.

Task: Find a rational number whose square is close to a given number.

Procedure: Let the given number be called n.

1. Make a guess of a number that is close to the square root of n, and call your guess a.
2. Compute n/a.
3. Find the average of a and n/a.
4. The number you just found is your new approximation of the square root of n. If you think it is good enough, you can stop. If you want to keep going, call this new number a and go back to step 2.

Let's see how that works with the numbers we used earlier.

Example: We want to find an approximation of the square root of 7200. We know that 8^2 is 64, so 80^2 is 6400. That is less than 7200 but not so far off, so we will make 80 to be our guess.

Now we need to compute the quotient 7200/80. This is 90. Find the average of 80 and 90, which is 85. That is the updated estimate. If we want we can stop here. We can also keep going. Let's keep going.

We now need to compute 7200/85. You can do this by hand. The quotient is not an integer. You can stop once you get a couple places past the decimal point. You should have roughly 84.70. Just cut off the end of the number, since we are making an estimate, and call it 84.7. Now we need to find the average of 85, the old estimate, and 84.7, the thing we had called n/a. You can check that the average of those two numbers is 84.85. This turns out to be a very good estimate. How do we know? Let's compute its square. The square 84.85^2 is 7199.5225. That is very close. The square of our estimate is off by less than half of a square unit. This concludes the example calculation.

Ordinarily, it is enough to go through the steps one or two times. In other words, you start with a reasonable guess and refine it once, possibly. We will not need accuracy beyond that.

23.2 *Finding Chords*

Sometimes we can find a chord directly, like we did above, using our knowledge of elementary geometry. At other times we need to proceed by an indirect route, finding the thing that we want by means of some intermediate calculations. We begin by computing directly with geometry using special figures.

23.2.1 Special Figures

We found the chord of 90° by using an isosceles right triangle, i.e., half of a square cut along its diagonal. We can use other special figures to determine other chords.

Proposition 104. *The chord of* 60° *is* 60.

Proof. Consider a regular hexagon circumscribed by a circle. The hexagon can be divided into six equilateral triangles, with the side of each triangle the same as the radii of the circumscribing circle. The diameter of the circle was taken to be 120 units, so the radius of 60 units, and the chord of 60° is also, therefore, 60 units. See Figure 23.2. □

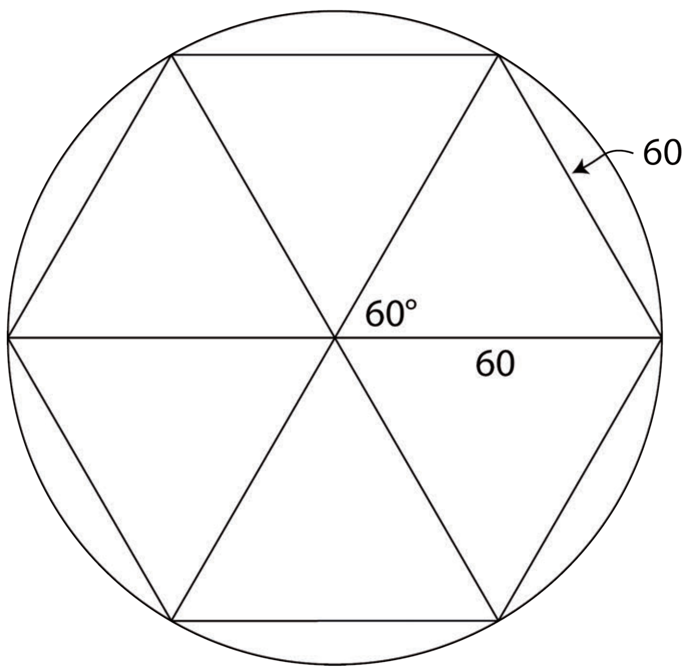


Figure 23.2: Chord and inscribed hexagon

It is important to realize that the relationship between angles and their corresponding chords is complex. The chord of 60° is very simple. We find that it is a specific whole number. For many angles this will not be the case. The angle 120°, twice the angle we just considered, is such an example.

Proposition 105. *The chord of* 120° *is approximately* 104.

Proof. Consider Figure 23.3. The large triangle, triangle BDE, is an equilateral triangle. We wish to determine the length of BD and will do so by finding the length of CD, which is half the total.

A smaller triangle, triangle ACD, is a right triangle (right angle at

The angle ADC is half the angle in an equilateral triangle, so it is 30°.

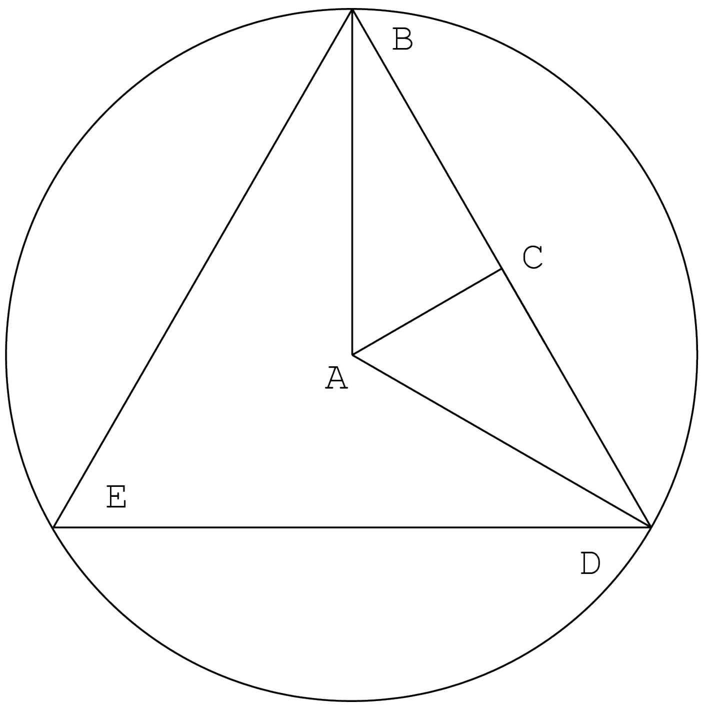


Figure 23.3: Chord of 120°

C) whose other angles are $60°$ and $30°$. This smaller triangle can in fact be considered as half of a different equilateral triangle, as seen in Figure 23.4. Since the radius of the circle is 60 units, the hypotenuse of the right triangle (i.e., AD) is also 60 units. The shorter side of the right triangle is 30 units, since it is half of the side of the (smaller) equilateral triangle in Figure 23.4.

We can now use the Pythagorean theorem to determine the remaining side. We know that the square on the hypotenuse is 3600 square units, and the square on the shorter leg is 900 square units. Thus, the difference, 2700 square units, is the square on the remaining leg. We now approximate the square root.

Let 50 be an initial guess. Carrying out the square root algorithm, we get the refined estimate of 52.

To complete the computation, observe that the longer leg of the small right triangle, segment CD, is half of BD, which is the chord determined by the angle BAD, measuring $120°$, that sits at the center of the circle in Figure 23.3. Doubling 52, we obtain 104. □

Recall that the golden ratio played an important role in our construction of a regular pentagon. We use now use our knowledge of ratios in the regular pentagon to determine a chord.

Proposition 106. *The chord of* $72°$ *is approximately* 71, *and the chord of* $144°$ *is roughly* 114.

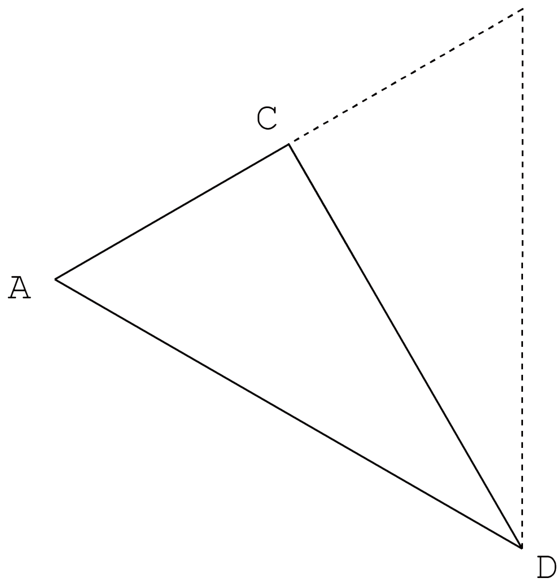


Figure 23.4: Another equilateral triangle

Proof. Consider the circumscribed regular pentagon, together with an isosceles triangle, in Figure 23.5. Recall that the ratio of the long side of the triangle to the side of the pentagon (which is also the short side of the triangle) is the golden ratio. We wish to find the length of one side of the pentagon, to obtain the chord of 72°, since 72 divides 360 into five parts.

Consider the labeled Figure 23.6. The point A is the center of the circle. The triangles BCA and BDE are right, and share the angle at B, thus they are similar triangles.

To show that the angle at C is right, reason about symmetry. What kind of triangle is ABE, for instance?

Similarity means that the ratio $AB : AC$ is the same as the ratio $BE : DE$. The latter ratio is twice the golden ratio, since it comes from the long side of the isosceles triangle and half the short side. You will show in an exercise that this ratio is approximately the ratio $32 : 10$, which we can write as the decimal number 3.2.

The side AB is 60 units, since it is a radius. Thus, we see that the ratio $60 : AC$ is approximately the same as $32 : 10$. Find the length AC by dividing 60 by 3.2. Doing this, we find that AC is approximately 18.7 units long.

This is like a very simple kind of algebra.

Use the Pythagorean theorem to determine the length of BC. The square on AB is 3600 square units, and the square on AC is approximately 350 square units (350 is roughly 18.7^2), so we can find BC by finding the root of the difference, which is 3250.

To find the root of 3250, begin with the estimate of 55 (since we know 60 is too big and 50 is too small). The quotient $3250/55$ is roughly 59. The average of 55 and 59 is 57. We conclude that BC is about 57 units long.

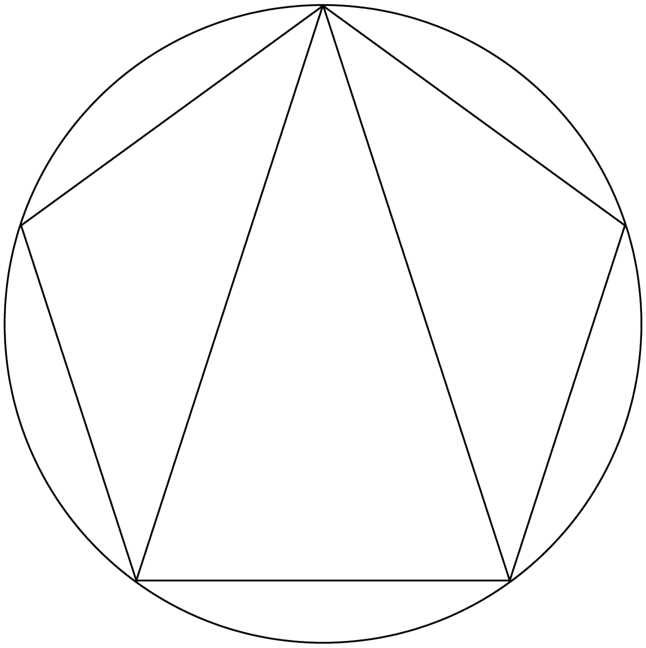


Figure 23.5: Isosceles triangle in pentagon

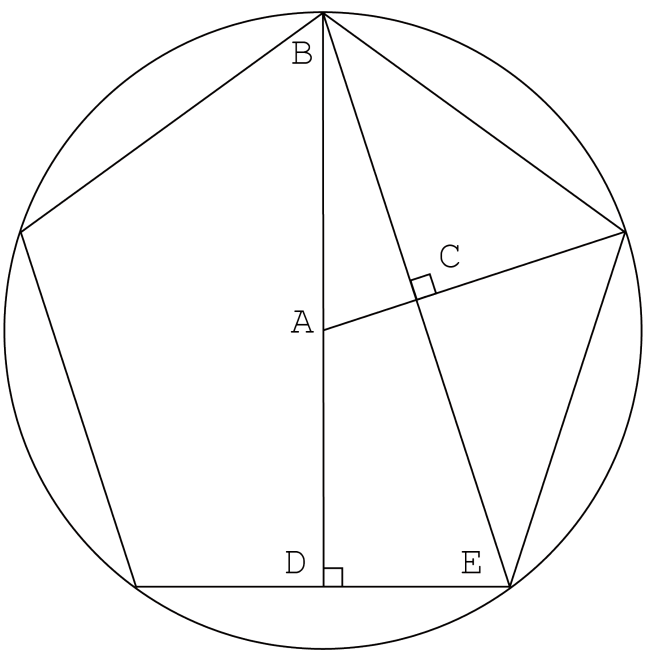


Figure 23.6: Right triangles in pentagon

Now find the length of DE using similarity. We know that the ratio of the shorter leg to the hypotenuse is the same in triangles BCA and BDE. Thus, $AC : BC$ is the same as $DE : BE$. We also see from symmetry that BE is twice BC, so that BE is roughly 114 units long. Thus, we have that the ratio $18.7 : 60$ is about the same as the ratio $DE : 114$. We conclude that DE is about 35.5 units in length. This means that the side of the pentagon, which is twice DE, is about 71 units in length.

We need not do any more work to find the chord of 144°. This is given by the segment BE, which we found to be 114 units in length.

□

The approximation 71 for the chord of 72° is in fact a slight overestimate. We will refine the calculation later.

23.2.2 General Principles

The calculations that we have done so far relied on some special features of specific shapes. In each case we came up with a specific argument using some particular geometrical characteristics. Now we want to find a methodical way to build on these calculations.

We will introduce two methods for finding chords of angles. The first is narrower, and the second is more general. You will show in an exercise that the first is in fact contained in the second. Even though the first way is narrower, we will begin with it, since it is easier to understand.

The first way allows us to find the chord of an angle when we know the chord of a different angle, the angle twice as large as the given one. Alternatively phrased, we can find the chord of a half-angle once we know the chord of a given one. This result is likely due to Archimedes. The situation is depicted Figure 23.7.

Proposition 107. *Let the angle BAD and its chord BD be given in the circle with center A and diameter BC, and let the ray AE bisect the angle ADB. Let EF be perpendicular to BC at F. Then the square on the segment BE is the same as the rectangle on BC and BF. The segment BF is half the difference between BC and CD.*

Before proving this proposition, let us discuss what it means. We are interested in finding the segment BE, since it is the chord of half the angle. According to the proposition, we are able to know the square of BE using other geometric information. Once we know the square, we can use our algorithm for root extraction to find the segment BE itself.

There is a second point to consider as well. How will we find BF? The segment BC is the diameter, so it is known at the outset; it does

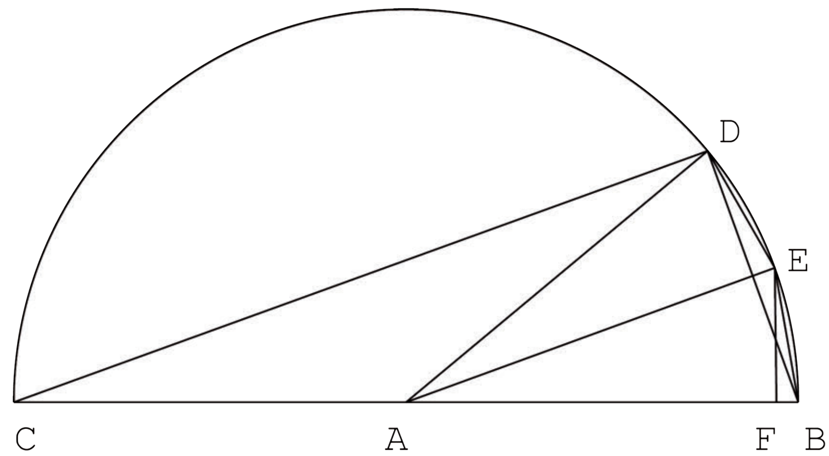


Figure 23.7: Chord of half angle

not depend on the angle. The segment BF, on the other hand, does depend on the angle BAD. This proposition, even though it reveals a real relationship among segments, will only be useful if BF can be found easily. The second conclusion of the proposition is that BF can be found through things that we already know or that can be computed without difficulty.

Proof. First, we prove that the square on BE is the same as the rectangle on BC and BF. Recall that, by Proposition 25 (III.31), a triangle in a circle with the circle's diameter as a side is right. Thus, the triangle CEB is right, with right angle at E. By Proposition 45 (VI.8) we know that triangle EFB is similar to triangle CEB. Thus, the ratio $BE : BF$, hypotenuse to shorter side, is the same as the ratio $CE : BE$.

In Chapter 8 of Geometry we saw that Proposition 53 (VI.14) allows us to infer equality of parallelograms from an equality of ratios of sides. By applying that proposition here, we conclude that the square on BE is the same as the rectangle on BC and BF, which was to be shown.

The second thing we need to do is to show that BF satisfies the relationship expressed in the proposition. To show this, introduce the point G on the diameter so that CG is the same as CD. See Figure 23.8.

You are capable of understanding Euclid's Book III, but we did not have the opportunity to study it. If you wish, you can look at it on your own in order to understand this proof in full detail. It is also fine in mathematics at times to accept and build upon results that other people prove. You do not have to do everything yourself.

It turns out that the angles DCE and GCE are equal. This is something that you can learn about if you study Book III of Euclid's *Elements*. We will see III.20 later (Proposition 129) and the result we are using now follows from that proposition.

Since CG and CD are the same, and CE is shared, and the two angles just mentioned are the same, we conclude that triangles CDE and CGE are the same. In particular, DE and EG are the same. Rely-

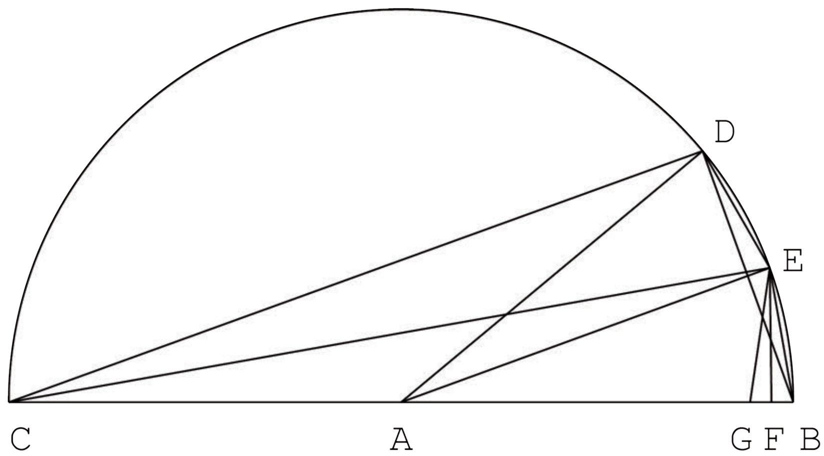


Figure 23.8: Half angle and auxiliary point

ing again on Euclid's Book III, DE is the same as BE, and therefore GE and BE are equal.

Since GE and BE are equal, F is the midpoint of BG. The segment BE is the difference of BC and CG which is the same as the difference between BC and CD. Thus, BF, which we wish to find, is half the difference between BC and CD. □

Observe that CD must be found. This can be done since the diameter BC and the segment BD, the chord of the original angle, are known. The segment CD combines with those other segments to form a right triangle, and so CD can be found using the Pythagorean theorem.

Here is an example of how the proposition can be used to find a chord.

Example: Find the chord of the angle $30°$.

We already know the chord of the angle $60°$, which is 60 units. Thus, we can find the chord of $30°$ using the procedure just given.

First, we find the segment that had been labeled BF. This is one half the difference between the diameter (120 units) and the segment that had been labeled CD. To find that segment, we need to find the root of the difference $120^2 - 60^2$. You can compute this yourself; it is approximately 104. Thus, the segment of interest is one half of $120 - 104$, so the segment (that was called BF) is about 8 units in length.

Next, the square of the chord we want to find is the same as the product of 8 and 120, which is 960. We need to find the square root of 960; it is about 31.

We will now consider a second way to find chords of angles using other known chords. This more general way seems to be due to Ptolemy, the author of the book *The Almagest*, which our current study of astronomy is based.

Theorem 108. *Let a quadrilateral inscribed in a circle be given. The rectangle on the diagonals is the same as the sum of the rectangles on the two pairs of opposite sides.*

You will complete the proof of this proposition as an exercise.

The consequence of this proposition is that if we know the chords of two angles, we can also find the chord of their sum and their difference. Here is an example of each kind of calculation.

Example (difference of angles): Find the chord of the angle 42°.

Consider Figure 23.9, in which angle BAC is 30° and angle BAD is 72°. It follows that angle CAD is 42°, so that the chord CD which corresponds to this angle is the desired quantity.

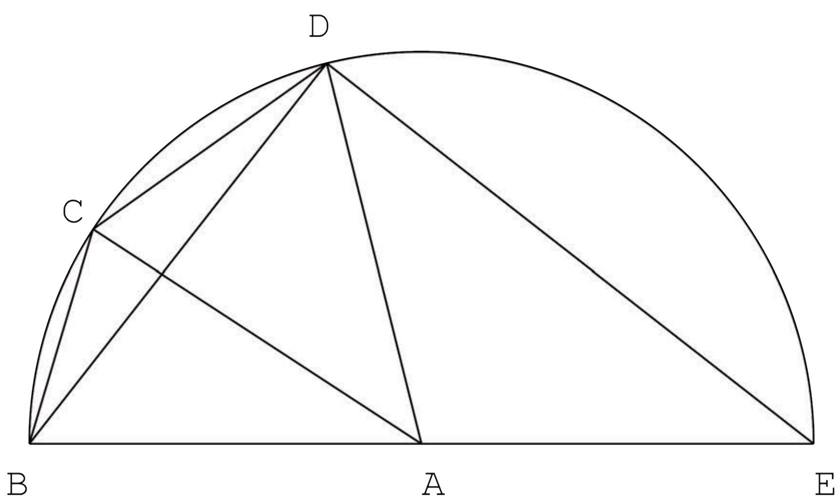


Figure 23.9: Chord of a difference

We know that the chord BC of 30° is roughly 31, and the chord BD of 72° is roughly 71. Use the Pythagorean theorem to find that DE is about 97. Use it again to find that CE is about 116. The segment BE, being the diameter, is 120. Then the theorem of Ptolemy says that the sum $31 \times 97 + CD \times 120$ is approximately the same as 71×116.

Recall that Proposition 25 (III.31) applies, telling us that a triangle formed by a diameter and a point on the circle is right

Since 31×97 is 3007, and 71×116 is 8236, we conclude that $120 \times CD$ is approximately 5229, so that CD is roughly 43.6.

Warning: We have made many approximations. This is necessary since we are dealing with quantities that cannot be expressed exactly as ratios of natural numbers. If we choose coarse approximations, and then use those results when determining other quantities, the new results will also tend to be coarse approximations. That is what has happened here. The chord of 42° is in fact very close to 43, but our estimate of 43.6 differs from this by more than 0.5 units, a fairly substantial error. In order to get a more accurate estimate of the chord of 42°, we would need to go back to compute the chords of 30° and 72° with greater precision. Then, using those more precise values, we would obtain a more precise value for the chord of 42°. You will have the chance to do this in an exercise.

Example (sum of angles): Find the chord of the angle 102°.

We can use once again use Ptolemy's theorem (Theorem 108) to answer this, since we know the chords of 72° and 30°, and so we can use those chords to find the chord of the sum of those angles.

Consider Figure 23.10. The angle BAC is 30° and the angle CAD is 72°. The chords BC and CD, which correspond to 30° and 72° respectively, are known. We wish to find BD, the chord of 102°. Introduce point F on the circle so that EF is the same as BC. This means that angle EAF is 30°.

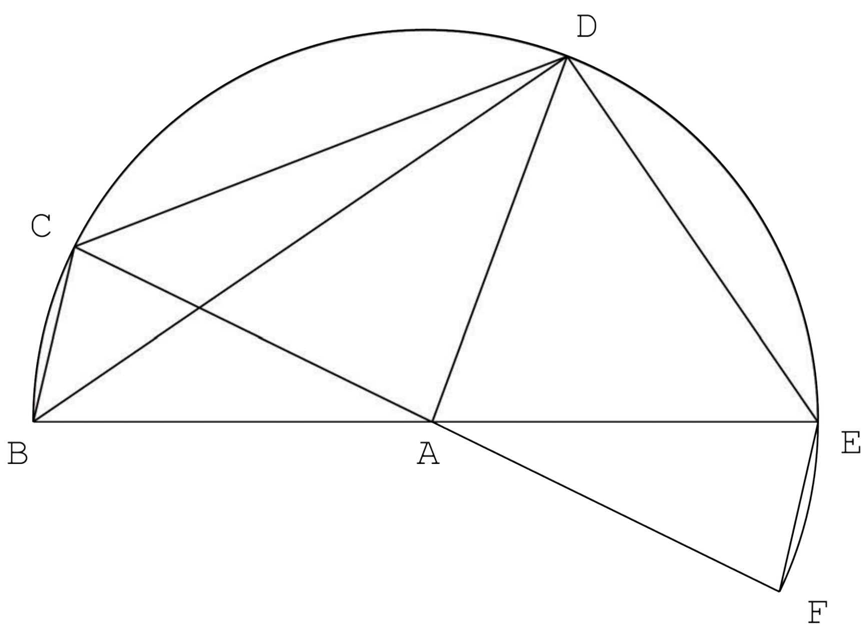


Figure 23.10: Chord of a sum

We will not find BD directly. Instead, we will first work with the quadrilateral $CDEF$ and find the length of DE. Since CD is known, we can use the Pythagorean theorem with the right triangle CDF to find DF. We need to approximate the root of $120^2 - 71^2$. This is roughly 97. Again, since EF is known, we can use the Pythagorean theorem to reason about the right triangle FEC. We see that EC is the

While we use the segment DF, it is not drawn in the figure given. This simply keeps the diagram uncluttered.

square root of $120^2 - 31^2$, which is roughly 116.

Now use the theorem of Ptolemy, Theorem 108, to find DE. We know that the sum $120 \times DE + 71 \times 31$ is about the same as the product 97×116. Thus, $120 \times DE$ is approximately 9051, so that DE is about 75.4.

To conclude, we can use the Pythagorean theorem and the right triangle BDE to find the desired chord, BD. We must approximate the square root of $120^2 - 75.4^2$. This is about 93.

23.3 Chord Table

At this point we have what might seem to be a fairly arbitrary collection of chords. Here they are in a table.

angle	chord
30	31
42	43.6
60	60
72	71
90	85
102	93
120	104
144	114

Remember that these are approximations, except for the special case of 60° where we can determine the chord exactly using an inscribed hexagon.

This table is too sparse to be of general use. What if we need to work with a 10° angle? Or an angle that measures 171°? We must refine the table.

It is possible to refine the table systematically. The key is that whenever we know the chord of two angles, we can find the chord of their sum and the chord of their difference, as we saw in the examples of 42° and 102°. We can also find chords of half angles. Using these principles, we slowly work our way out from things that are known to things that are unknown to us.

Here is the strategy we will use to get an accurate chord table.

1. Find the chord of 12° using the angles 60° and 72°.

2. Use the half-angle theorem to find the chord of 6°.

3. Use the half-angle theorem again to find the chord of 3°.

4. Use the methods for finding chords of sums and differences to build in 3° increments from known values.

The final step will be clearer once we get there. The result of our work will be to have a reasonably precise table of chords of each angle, in 3° increments, between 3° and 180°.

23.3.1 *Chord of* 12°

Before we compute the chord of 12°, we need to improve a prior calculation. We know that the chord of 60° is exactly 60. There is nothing that can be improved in that number.

The chord of 72°, on the other hand, was computed without a great deal of precision. We used a regular pentagon within a circle, and used the approximation $32 : 10$ to approximate the ratio that is twice the golden ratio.

We will now go through the same steps from that proposition, Proposition 106, but taking care to have greater accuracy at each step. It turns out that a better estimate than 3.2 is possible. The details are given in an exercise. Rather than taking 3.2, we will use the value 3.236. Proceeding as before, the quotient $60/3.236$ is approximately 18.541.

Our coarser estimate had this as 18.7.

The square of 18.541 is about 343.769. We then need to find the root of the difference $3600 - 343.769$. This is about 57.06.

Finally, use similarity as before. Half the quantity we seek is $114.12 \times 18.541/60$. This is about 35.265. Doubling that, we obtain the approximation 70.53 for the chord of 72°.

Now, we use the theorem of Ptolemy to find the chord of the difference $72 - 60$. This is done just like the example for finding the chord of 42. We use the refined approximation 103.92 for the chord of 120° (using the Pythagorean theorem). We use the refined approximation 97.09 for the chord of 108° (the other part of the triangle made by the diameter and the chord of 72°). The result is that the product of 120 and the quantity we seek is the difference $70.53 \times 103.92 - 60 \times 97.09$, about 1504.1. Dividing by 120, we find that the chord of 12 is about 12.53.

In fact 12.54 is a better approximation than 12.53, but we would need greater precision in earlier steps to obtain this.

23.3.2 *Chords of* 6° *and* 3°

We know that the chord of 12° is about 12.53. We use the half-angle chord computation, Proposition 107, to find the chord of 6°.

The root of $120^2 - 12.53^2$ is roughly 119.34. We then must find half the difference between this and the diameter, which is 0.33. Finally, we must find the root of 120×0.33. This is about 6.29, which is thus our estimate of the chord of 6°.

We now follow the same procedure to approximate the chord of 3°. The root of $120^2 - 6.29^2$ is roughly 119.835. The difference between this and the diameter is just 0.165, half of which is 0.0825.

There are some careful choices that are hidden here. They help us arrive at a better approximation.

The root of 120×0.0825 is about 3.15, giving us a value for the chord of $3°$.

It turns out that, with the approximations we started with, we have ended up with slight overestimates. That is fine. We are within about a hundredth part in our values for the chord of $6°$ and the chord of $3°$.

23.3.3 Sources of Inaccuracy

Note: *This is an optional section for those who wish to analyze computations more carefully.*

There is a feature of the half-angle computational method worth noting. We considered expressions of the form $120^2 - (\text{small})^2$, where "small" refers to the fact that we are using a fairly small number there. The numbers we obtain in this way will have square roots that are very close to 120.

We then found the difference of 120 and that other number, which was very close to it. The result is that we get a very small quantity, since the two hardly differ. This is a delicate matter when we seek to do calculations with a high degree of accuracy. If we have two things that we know fairly well, and they are close together, and then we have to use their difference, this can be a source of significant deviation.

Let's make this concrete with an example. In the calculation of the chord of $3°$ we chose 119.835 as our estimate of the relevant root. Then the difference $120 - 119.835$ is 0.165, half of which is 0.0825. Multiplying by 120, we found that we needed to compute the square root of 9.9.

Suppose that we had made a slightly different choice of approximation, where the root we start out with is taken to be 119.83. This is different than the previous choice by only 0.005. Then the relevant difference is 0.17, half of which is 0.085. Multiplying by 120, we get 10.2, so that we would approximate the root of 10.2.

The square root of 10.2 is roughly 3.19. This is fairly different from our chord estimate of 3.15. The small variation (five thousandths) at one point in the calculation led to a fairly large change (four hundredths) in the output. The variation in the early estimates yielded variation in the conclusion almost ten times as large. The reason is that we needed to use the difference of two quantities that are very close to each other.

The chord of $3°$ is in fact close to 3.141, so we see that 3.15 is a better estimate than 3.19.

23.3.4 The Table in $3°$ Increments

With a good approximation of the chord of $3°$ we can fill out a table step by step. Here is a sketch of how the entries can be systematically found.

- To find the chord of 9°, use the fact that 9 is the sum of 3 and 6.
- To find the chord of 51°, use the fact that 51 is the difference of 60 and 9.
- To find the chord of 171°, use the fact that the sum of 171 and 9 is 180 (i.e., use the Pythagorean theorem for a triangle on a diameter in a circle.) More generally, once the chord of an angle α is known, the Pythagorean theorem allows us to find the chord of $180 - \alpha$.
- Carry out earlier computations with greater precision to get better values for angles like 90° and 30°.

angle	chord	angle	chord	angle	chord
3	3.15	63	62.7	123	105.46
6	6.29	66	65.36	126	106.92
9	9.42	69	67.97	129	108.31
12	12.53	72	70.53	132	109.63
15	15.66	75	73.05	135	110.87
18	18.77	78	75.52	138	112.03
21	21.87	81	77.93	141	113.12
24	24.95	84	80.3	144	114.13
27	28.01	87	82.6	147	115.06
30	31.06	90	84.85	150	115.91
33	34.08	93	87.04	153	116.68
36	37.08	96	89.18	156	117.38
39	40.06	99	91.25	159	117.99
42	43.01	102	93.26	162	118.52
45	45.92	105	95.2	165	118.97
48	48.81	108	97.08	168	119.34
51	51.66	111	98.9	171	119.63
54	54.48	114	100.64	174	119.84
57	57.26	117	102.32	177	119.96
60	60	120	103.92	180	120

Figure 23.11: Chord table

23.4 *Menelaus in the Plane*

We will now examine a property of configurations of lines in the plane. The natural way to phrase this property is as an equivalence of ratios. First, though, we must consider the notion of compounding of ratios.

23.4.1 *Compounding Ratios*

It is likely that you are familiar with doing arithmetic with fractions. This includes multiplication of two fractions. Earlier in the book, when we first studied ratios, we did so very carefully, thinking in terms of repetitions. At this point we need to abandon that careful treatment and use the more familiar notion of multiplication of fractions.

Suppose that we have the ratios $a : b$ and $c : d$ where a, b, c, and d are numerical things. Then the ratio compounded of these two ratios is $ac : bd$. The way that we denote compound ratios symbolically is $(a : b).(c : d)$. At this point you can write the ratios as fractions if you wish, with the symbol / for division rather than :, the colon used for Eudoxan ratio.

23.4.2 *Properly Geometric Compounding*

This is an optional section for advanced students.

In some cases, we can understand compound ratio purely geometrically, without recourse to numbers. Suppose that a, b, c, and d are all segments. Then the ratios $a : b$ and $c : d$ exist. It is possible to construct a notion of the compound ratio $(a : b).(c : d)$ without introducing any units of measurement.

Let $a.c$ denote the rectangle formed by a and c, and let $b.d$ denote the rectangle formed by b and d. The two rectangles have a ratio, $a.c : b.d$, and this ratio of rectangles can be given as the definition of the compound ratio of the segments.

You can investigate how to compound the ratio of two rectangles and two line segments by a single ratio of solids. Beyond this it is difficult to give a direct interpretation in elementary geometry.

23.4.3 *Menelaus and Ratios in Line Segments*

Proposition 109. *Let segments be given as in Figure 23.12. Then the ratio $AC : AB$ is the same as the compound ratio $(CE : DE).(DF : BF)$.*

Before proving the proposition let's see how to remember it. The ratio $AC : BC$ involves a whole and a part. The ratios $DF : BF$ and $CE : DE$ involve wholes and parts, in the opposite order. Finally, the parts DE and DF are both remote from the segment AC, i.e., they do not touch it. So a way to remember the content of this proposition, if you do not remember a specific labeling of the points, is that it says that a ratio of a whole to a part is the same as the ratio of a whole to a remote part compounded with the ratio of a remote part to a whole. ("Remote" is said here with respect AC, the first "whole" in question.)

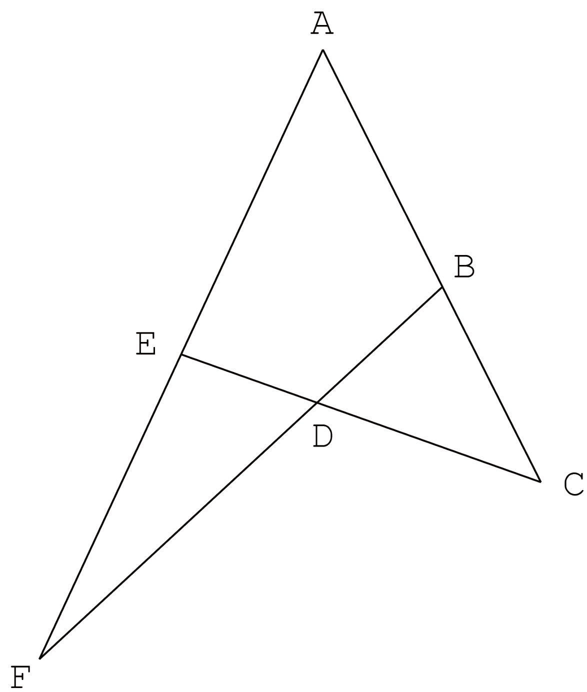


Figure 23.12: Ratios of line segments

Proof. Consider Figure 23.13 in which the line BG parallel to CE has been added. Then the triangles ABG and ACE are similar, and the triangles DEF and BGF are similar. By the first similarity, the ratios $AC : AB$ and $CE : BG$ are the same. By the second similarity, the ratios $DE : BG$ and $DF : BF$ are the same.

We now use a trick. This next step is not obvious. Its utility will be revealed as we continue.

The ratio $CE : BG$ is the compound of $CE : DE$ and $DE : BG$. Note that DE is the second term in the first ratio and the first term in the second ratio.

Since $DE : BG$ and $DF : BF$ are the same, we see that $CE : BG$ is the compound ratio $(CE : DE).(DF : BF)$. Since $CE : BG$ is the same as $AC : AB$, we have obtained the conclusion that we sought.

□

You can think of the tricky step in terms of fractions; we have introduced the same term in the numerator and denominator simultaneously. If you looked at the advanced section on the geometric interpretation of compound ratio, the equality of ratios can be shown using the geometric definition of compound ratio along with Proposition 42 (VI.1).

There is a similar statement about the ratio of a part to a part, rather than a whole to a part, in a similar figure. We only state the proposition here. Its proof is an exercise.

Proposition 110. *Let segments be given as in Figure 23.12. Then the ratio $AB : BC$ is the same as the compound ratio $(DE : CD).(AF : EF)$.*

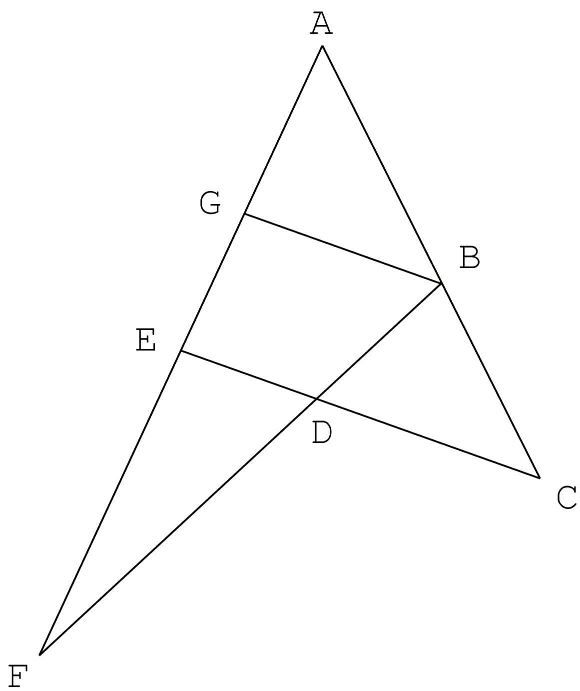


Figure 23.13: Ratios of segments with auxiliary point

23.5 *Menelaus on the Sphere*

The results of the last section can be used to reason about ratios arising from arcs on spheres. This kind of reasoning is useful when we want to consider motion in the sky. In order to translate the results from the plane to the sphere, we first need to address some preliminary technical matters.

23.5.1 *Preliminaries*

Two points on a circle determine an arc. That arc is the section of the circle that lies between them. Given points A and B on a circle, we will write

$$\text{arc } AB$$

Provided that the two points do not lie on a diameter, take the shorter section. Then there is a unique are. The case of the diameter is ambiguous.

to indicate the arc determined by these two points. The arc also determines an angle. If C is the center of the circle, the angle ACB is the angle corresponding to the arc that we call arc AB.

It is also convenient to have a simple way to refer to the chord of an angle, so that we do not need to write the entire phrase "the chord of the angle ..." each time. Given an angle ACB, we write

$$\text{crd } ACB$$

to indicate the chord of the angle. Recall that this is a measurement in which the diameter is taken as having 120 parts.

It will be useful to combine these two abbreviations. Two points A and B on a circle determine an arc, which corresponds to an angle at the center of the circle. We can consider the chord of that angle. The brief way to write this is as follows.

$$\text{crd arc } AB$$

In doing this, we rely on the fact that an arc and an angle can be identified in a natural way.

The next proposition is illustrated by Figure 23.14.

Proposition 111. *Let points A, B, and C be given on a circle with center D, with arcs AB and BC each less than a semicircle. Let E be the point at which the line through A and C intersects the radius that terminates in B. Then the ratio crd 2 arc AB : crd 2 arc BC is the same as the ratio $AE : CE$.*

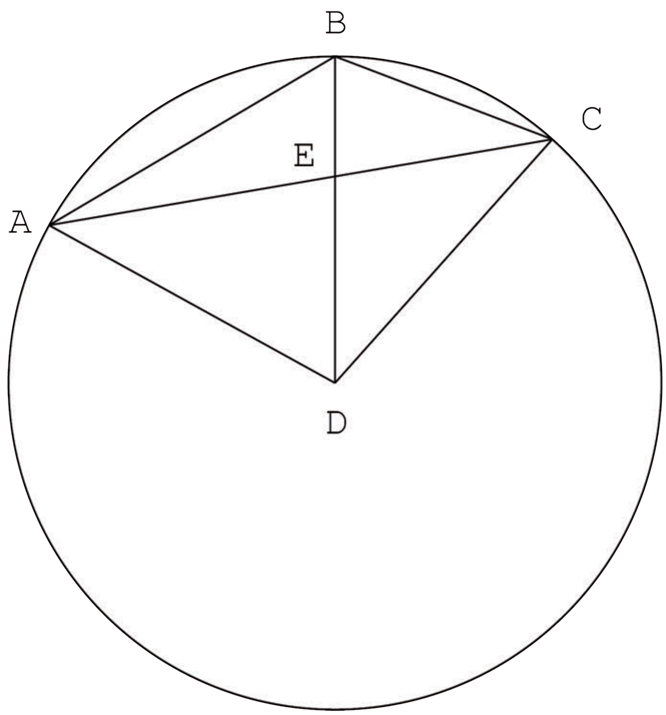


Figure 23.14: Ratios of chords

Before giving the proof it is important to clarify notation. The expression crd 2 arc AB means this. The points A and B determine an arc of a circle. That arc can be copied again immediately adjacent to itself to yield a new arc twice as large. That arc is called 2 arc AB. If you think instead using angles at the center of the circle, then 2 arc AB is the angle that is twice the angle determined by the arc AB.

Look back to the chords of 60° and 120° to see that crd 2 arc AB and 2 crd arc AB are different. One is "the chord of twice the arc," while the other is "twice the chord of the arc."

Proof. Let F be the point on the radius DB such that the line AF is perpendicular to that radius, and let G be the point on the radius DB so that CE is also perpendicular to DB. See Figure 23.15.

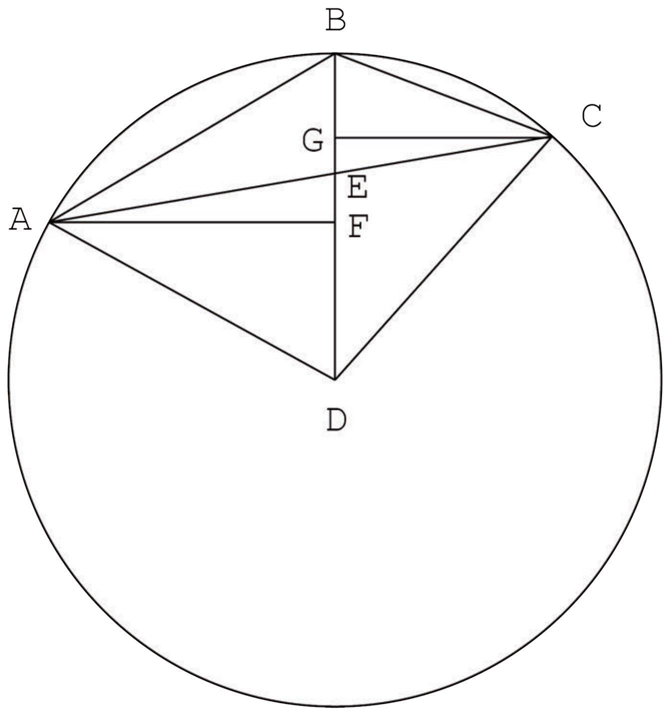


Figure 23.15: Ratios of chords

The angles AEF and GEC are vertical, thus equal, so the right triangles AEF and GEC are similar. This means that the ratio $AE : CE$ is the same as the ratio $AF : CG$. Observe that AF is one half of crd 2 arc AB, and similarly CG is one half of crd 2 arc BC. (See Figure 23.16.) Thus, the ratio $AF : CG$ is the same as the ratio crd 2 arc AB : crd 2 arc BC. Since $AE : CE$ is the same as $AF : CG$, the proof is complete.

Can you find the relevant Euclidean proposition about vertical angles in Part I?

□

A similar statement about ratios of chords of doubled arcs holds when we consider the line determined by two adjacent points among three on a circle.

The proof of the next proposition is similar to that of the previous one. It is left as an exercise. The situation is depicted in Figure 23.17.

Proposition 112. *Let two points A and C be chosen on a circle with center D, such that AC is not a diameter. Let a point B be chosen on the arc AC. Let E be the point at which the line through B and C intersects the line AD. Then the ratio crd 2 arc AC : crd 2 arc AB is the same as the ratio $CE : BE$.*

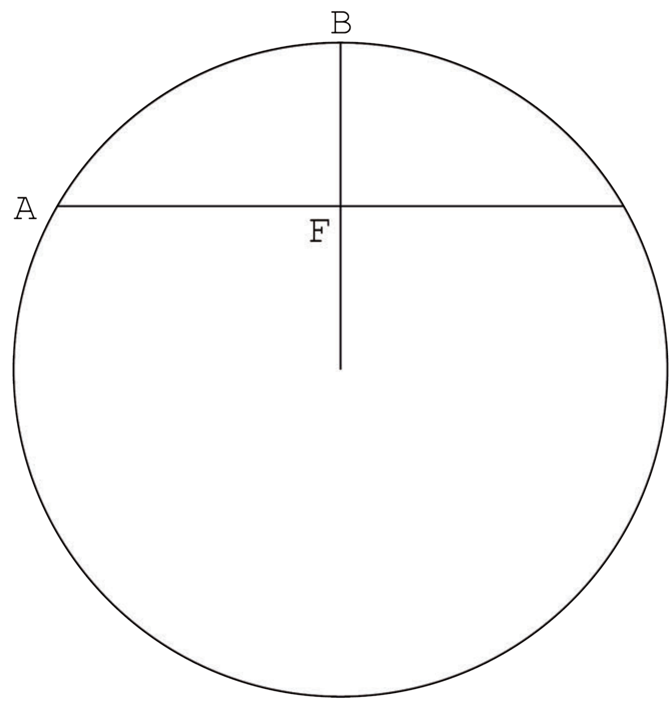


Figure 23.16: Double the arc

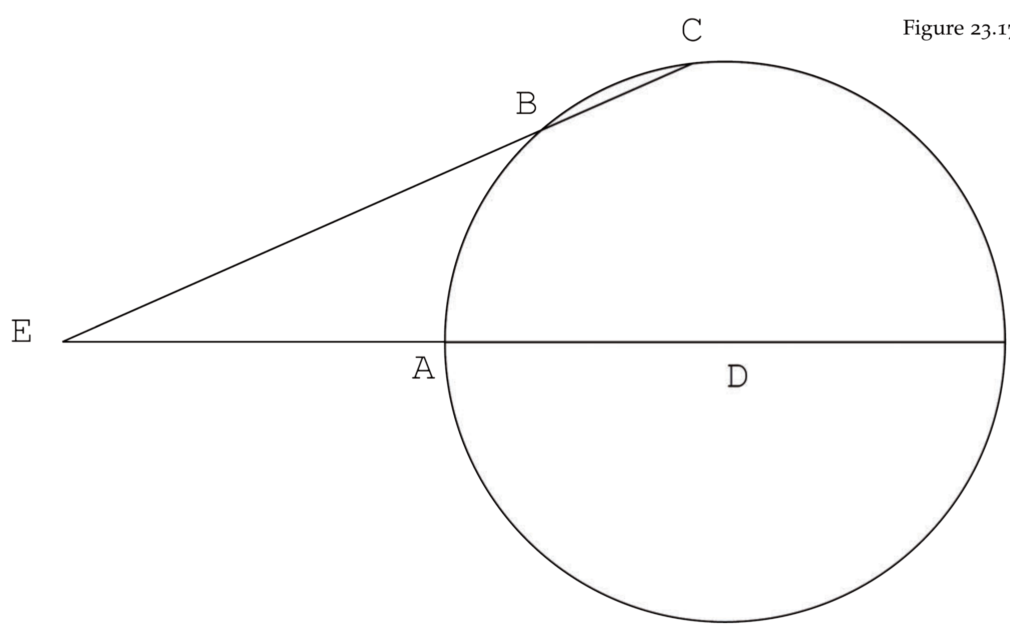


Figure 23.17: Angle outside circle

23.5.2 *Two Forms of Menelaus's Theorem*

Now we put our work with ratios to use in reasoning about arcs on spheres. The two results here, named for Menelaus, allow us to use our abstract knowledge of angles and chords to reason about the sun and moon.

Theorem 113 (Menelaus's Theorem). *Let a sphere with center O be given. Let points A and C on the sphere be given, and let B be a point between them on the arc of their great circle. Let another point F be given, and let E be a point between A and F on the arc of their great circle. Let the arcs CE and BF intersect in the point D. Then the following two statements hold.*

MT I: *The ratio crd* 2 *arc* AC : *crd* 2 *arc* AB *is the same as the compound ratio* (crd 2 *arc* CE : *crd* 2 *arc* DE).(crd 2 *arc* DF : *crd* 2 *arc* BF).

MT II: *The ratio crd* 2 *arc* AB : *crd* 2 *arc* BC *is the same as the compound ratio* (crd 2 *arc* DE : *crd* 2 *arc* CD).(crd 2 *arc* AF : *crd* 2 *arc* EF).

See Figure 23.18 for a depiction of the arcs on the surface of a sphere.

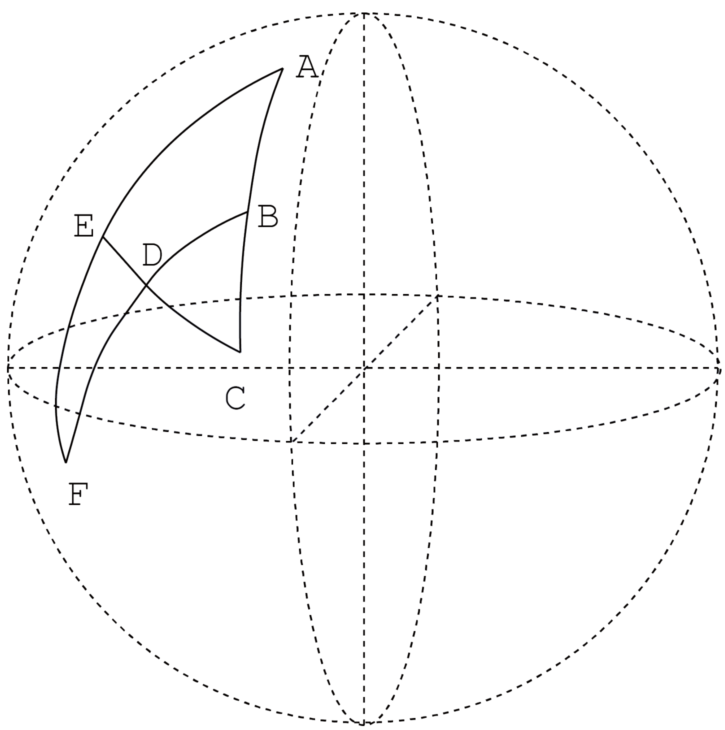


Figure 23.18: Menelaus's Theorem

You can see that the two forms of Menelaus's Theorem are quite similar to Propositions 109 and 110 we saw earlier about ratios of segments in triangular configurations in the plane. We will now give

The points in the statement of Menelaus's Theorem are named so that they will be in the same places as they were in the statements of the planar propositions.

a proof of the second part of Menelaus's Theorem, the part labeled MT II.

Proof. Consider the plane that contains O, A, and F. Within that plane, produce the line AE. Produce as well the line OF, and let the two lines intersect at point G. See Figure 23.19. By Proposition 112, the ratio crd 2 arc AF : crd 2 arc EF is the same as the ratio $AG : EG$.

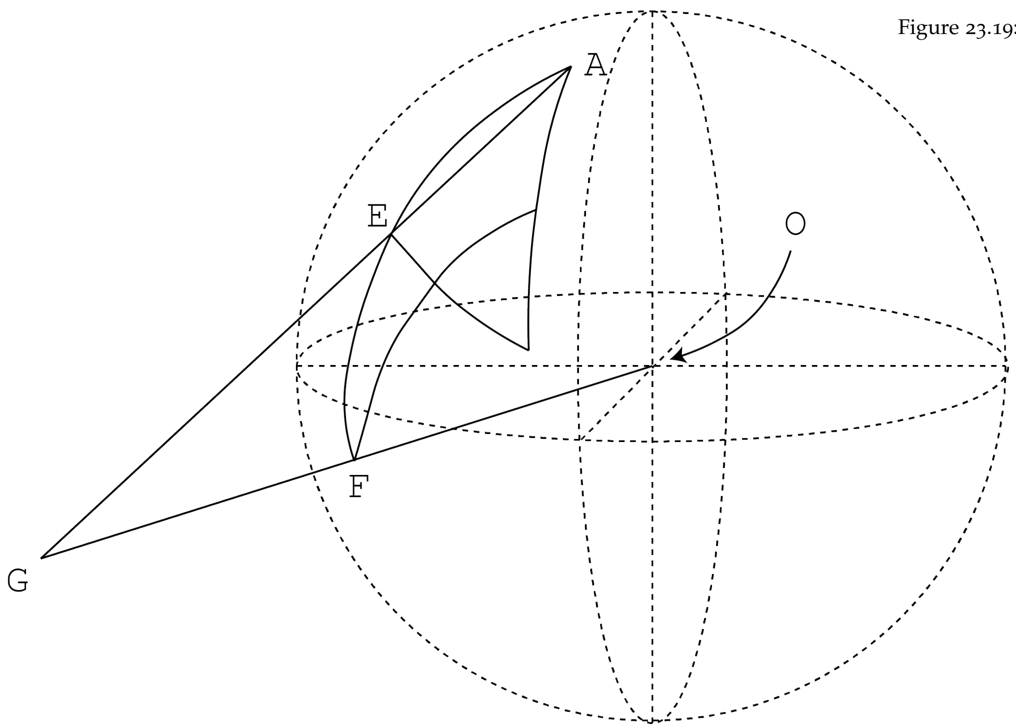


Figure 23.19: Menelaus's Theorem

Consider the plane that contains O, A, and C. Produce the line AC, and let H be the point at which AC intersects the radius OB of the sphere. See Figure 23.20. By Proposition 111, the ratio crd 2 arc AB : crd 2 arc BC is the same as the ratio $AH : CH$.

Produce the line GH. This line in fact intersects the sphere's radius OD. Let the point of intersection be K, as in Figure 23.21. You will show in an exercise that, reasoning similarly to previous cases, the ratio $KE : CK$ is the same as the ratio crd 2 arc DE : crd 2 arc CD.

Consider the plane containing the points A, C, and G. Note that it includes the chords AC and AE. Within that plane we have a configuration as in Figure 23.22. Use Proposition 110, regarding ratios of a part to a part in such a configuration. We see that the ratio $AH : CH$ is the same as the compound ratio $(KE : CK).(AG : EG)$.

The earlier paragraphs of the proofs showed that each of the ratios $AH : CH$, $KE : CK$, and $AG : EG$ is the same as a ratio of (doubled)

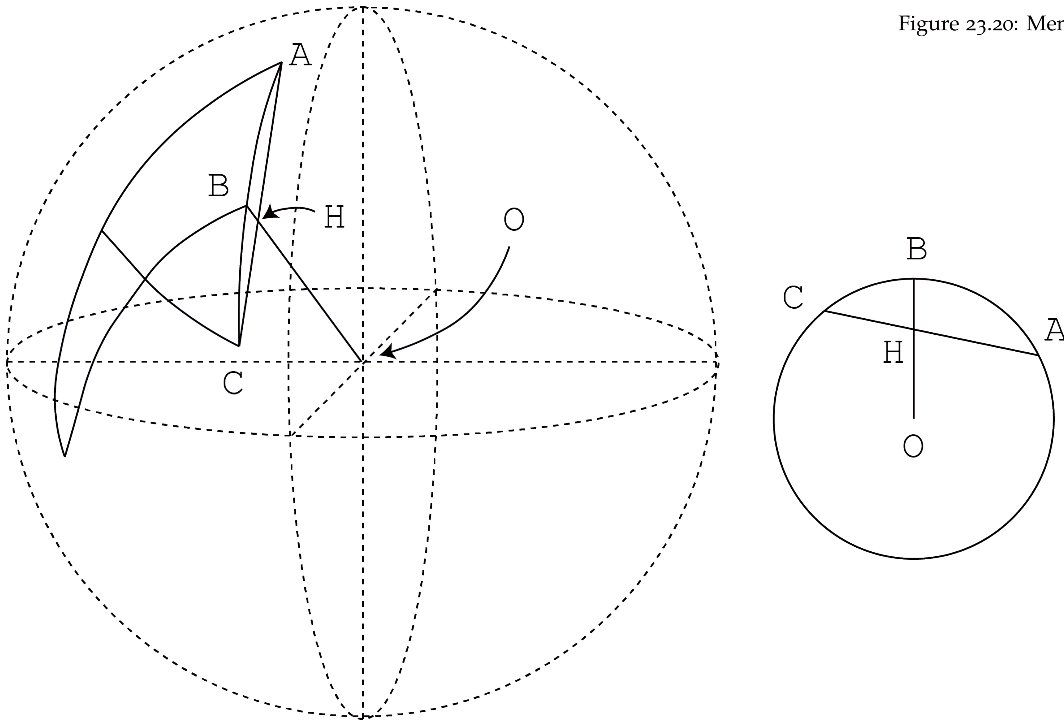


Figure 23.20: Menelaus's Theorem

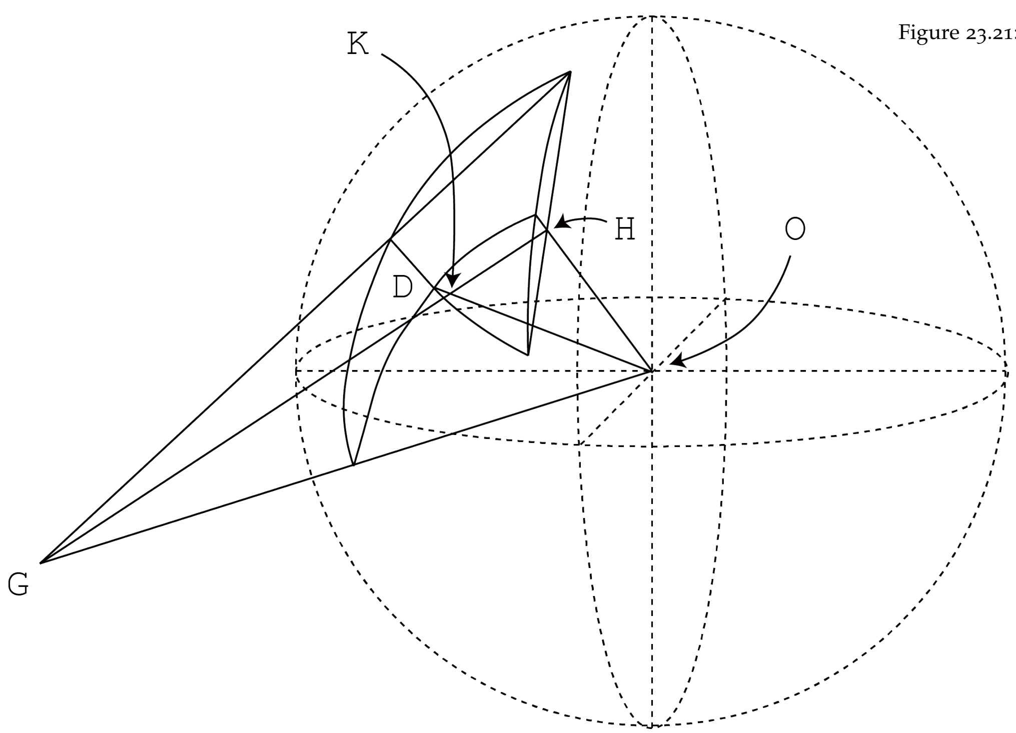


Figure 23.21: Menelaus's Theorem

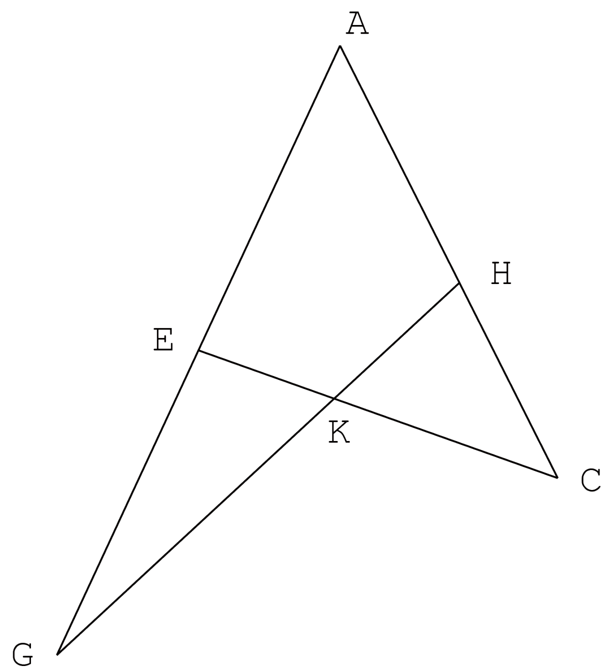


Figure 23.22: Menelaus's Theorem

chords. By restating the equivalence of ratios using the chord ratios rather than the segment ratios, we obtain the conclusion.

□

23.6 *Exercises*

Exercise 1

Try the algorithm for finding the square root of 7200, but start with 82 as your initial guess rather than 80. Compare this to the example that was given. You can stop after one iteration, or you can carry out a second one as well.

Exercise 2

Show that the ratio that is twice the golden ratio is roughly 3.2 using the following steps.

1. Produce a segment AB and divide it in the golden ratio at C. Recall that this means that the square on AC is the same as the rectangle on AB and BC.

Go back to page 69 for the procedure.

2. Produce a long line, much longer than AB (about 6 times as long).

3. On the line you have produced, copy segment AC eight times.

4. On the line, copy segment AB five times.

5. See that five copies of AB are nearly the same as eight copies of AC. Conclude that $AB : AC$ is close to the ratio 8 : 5

6. Doubling, we see that the ratio $2AB : AC$ is roughly 16 : 5.

7. Represent the ratio 16 : 5 as a decimal number (use long division).

Exercise 3

Explain the refined calculation of the chord of 72° using the following quantities. You do not need to find any numbers. Instead, you simply need to go through the steps of geometric reasoning shown earlier. Each time that you need an approximating number, it is available below. You need only to find the relevant one.

1. the product 18.541×3.236 is about the same as 60

2. the product 114.12×18.541 is about the same as 2115.9

3. $3600 - 343.769$ is the same as 3256.231

4. the product 35.265×60 is the same as 2115.9

5. 57.06^2 is about 3256

6. the product 57.06×2 is 114.12

7. 18.541^2 is roughly 343.769

8. the product 35.265×2 is the same as 70.53

Exercise 4

Choose three of the assertions about numbers given in the preceding exercise. Verify them, completing all the arithmetic by hand.

Exercise 5

Refine the calculation of the chord of 30° using the following quantities. You do not need to find any numbers. Instead, you simply need to go through the steps of geometric reasoning shown earlier. Each time that you need an approximating number, it is available below. You need only to find the relevant one.

1. the difference $120^2 - 60^2$ is the same as 10800

2. the square of 103.92 is about the same as 10799.4

3. the difference $120 - 103.92$ is the same as 16.08

4. twice 8.04 is the same as 16.08

5. the product 120×8.04 is the same as 964.8

6. the square of 31.06 is approximately 964.72

Exercise 6

Choose three of the assertions about numbers in the preceding exercise and verify them, completing all arithmetic by hand.

Exercise 7

Use Ptolemy's theorem and the following computations to obtain a better approximation of the chord of 42°. You do not need to find any numbers. You only need to reason using the ones that are given.

1. the chord of 72° is about 70.53
2. the chord of 30° is about 31.06
3. the square of 31.06 is about 964.72
4. the difference $14400 - 964.72$ is the same as 13435.28
5. the square of 70.53 is about 4975.48
6. the difference $14400 - 4975.48$ is the same as 9425.52
7. the square of 97.09 is approximately 9426
8. the product 31.06×97.09 is roughly 3015.6
9. the square of 115.9 is about 13432.83
10. the product 70.53×115.9 is about 8174.423
11. the difference $8174.42 - 3015.6$ is the same as 5158.82
12. the product 43×120 is the same as 5160

Exercise 8

Choose three of the assertions about numbers in the preceding exercise (other than the first two, about chords) and verify them by hand.

Exercise 9

The two preceding exercises showed how you produce an approximation of the chord of the angle 42°. The approximation you obtain is not the same as the one shown in the chord table, Figure 23.11. Discuss this seeming discrepancy. What are sources of error?

Exercise 10

Prove Proposition 25 (III.31) again.

Exercise 11

Copy the proof of Proposition 45 (VI.8), and then reread the proof of Proposition 107, which enables us to calculate chords of half angles. Observe the role that VI.8 plays in the latter proof.

Exercise 12

Find the chord of $15°$ using the half angle method, taking 31 as an approximation of the chord of $30°$.

Exercise 13

When we computed the chord of $30°$ using Proposition 107, we used the Pythagorean theorem to find the length of a chord. Examine that example again, and see how you can use the work there to find the chord of $120°$. Note that you do not need to compute anything. You simply need to interpret the numbers that are already given.

Exercise 14

Here is an outline of the proof of Ptolemy's theorem, Theorem 108. Fill in the details. It relies on thinking about compound ratios geometrically, as was explained briefly in Section 23.4.2.

Proof. 1. Let a circle be given, and let points A, B, C, and D be chosen (in order) on the circle, forming a quadrilateral $ABCD$. (Give a diagram illustrating this. Don't make them all the same distance apart. In other words, don't make a square.).

2. Let the point E be chosen on diagonal AC so that angle ADE is the same as angle BDC. (Add this to your diagram.)

3. The angles ADB and EDC are the same. (Explain why.)

4. Angles DBA and DCE are the same. (This follows from Proposition 129 [III.20] which you will see later. You can look ahead to that in order to justify fully the current claim, or you can just take it as given.)

5. Triangles ADB and DCE are similar. (Justify this by using the two preceding statements about angles, and the fact that two angles in a triangle determine the third [Proposition 24 (I.32)].)

6. The ratio $CD : CE$ is the same as the ratio $DB : BA$. (Use similarity.)

7. The rectangle on CD and AB is the same as the rectangle on CE and DB. (Think about rectangles in geometric compounding along with the notion of "clearing denominators" when ratios are thought of as fractions. In other words, compound both the

ratio $CD : CE$ and $DB : BA$ with the ratio $CE : CD$. By a different route, you can fully justify this by using Proposition 53 (VI.14).)

8. Show that triangles ADE and DCB are similar, and conclude that the rectangle on DA and BC is the same as the rectangle on DB and AE. (Reason as you did for the previous three items.)

9. The sum of the rectangle on DB and AE with the rectangle on DB and CE is the same as the rectangle on DB and AC. (Look at your diagram. How are AE, CE, and AC related?)

10. The conclusion of Ptolemy's theorem follows. (You added two rectangles in the preceding step. What are each of those rectangles equal to, individually, as determined in earlier steps?)

□

Exercise 15

Prove Proposition 110, the second part of the planar form of Menelaus's Theorem.

Exercise 16

Prove Proposition 112 about ratios of doubled chords.

Exercise 17

Proposition 112 about ratios of chords of doubled arcs presumes that the two lines intersect. Show that there are configurations of points A, B, C so that the lines AD and BC do not intersect. Refine the hypothesis of the proposition accordingly.

Exercise 18

In the proof of Theorem 113 (in which we proved the second part, MT II), there is one instance in which two ratios are asserted to be the same, though without proof. Prove that equality of ratios, using Proposition 111.

Exercise 19

Prove the first part of Menelaus's Theorem, MT I.

24 Principal Solar Events

We will now witness the extraordinary power of the theorem of Menelaus and our computation of chords.

24.1 The Solstices and Equinoxes

The summer solstice is the longest day of the summer. More precisely, thinking in terms of the celestial sphere, it is the point on the ecliptic that is furthest from the celestial equator, towards the north. When the sun is at that point, we have the longest day. Similarly, the winter solstice is the point on the ecliptic furthest from the equator to the south, and the day is shortest when the sun is at that position.

The equinoxes, vernal (spring) and autumnal (fall), are the points where the ecliptic and the equator intersect. When the sun is at one of those points, the day and night are of equal length.

This equality is approximate.

The ecliptic is at an angle to the celestial equator. This angle is called the "obliquity of the ecliptic." The angle is roughly 23.5°.

24.1.1 Length of Day

How long is the longest summer day at the latitude 39°? We can answer this question using our geometrical tools.

On the longest day of summer, the sun is at the summer solstice (considered as a point on the ecliptic). In Figure 24.1 you see a depiction of the sun at the horizon at dawn, along with the great circles of the celestial equator and the ecliptic.

A more abstract view is given in Figure 24.2. The point S denotes the location of the sun at the summer solstice point of the ecliptic. The line SA is the horizon. The points A and B are on the celestial equator, with A on the horizon. The point B is the point on the equator which will be at its highest point at noon, when the sun is at the meridian. The whole arc BS will be on the meridian at noon. The point E is the intersection of the equator with the meridian. The arc

Take time to see the connection between Figure 24.1 and Figure 24.2. You must be able to connect the abstract mathematical diagram to ordinary things that you can see, like the sun and the horizon.

EM continues below the horizon to P, the celestial south pole.

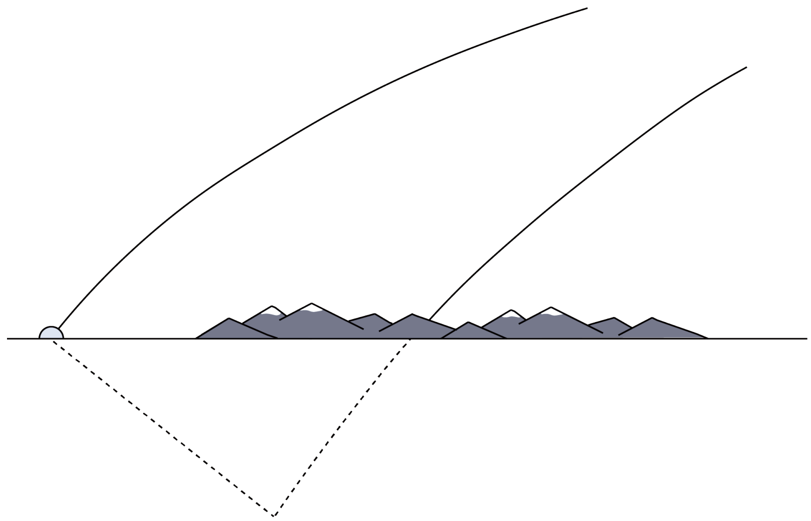


Figure 24.1: Sunrise at summer solstice

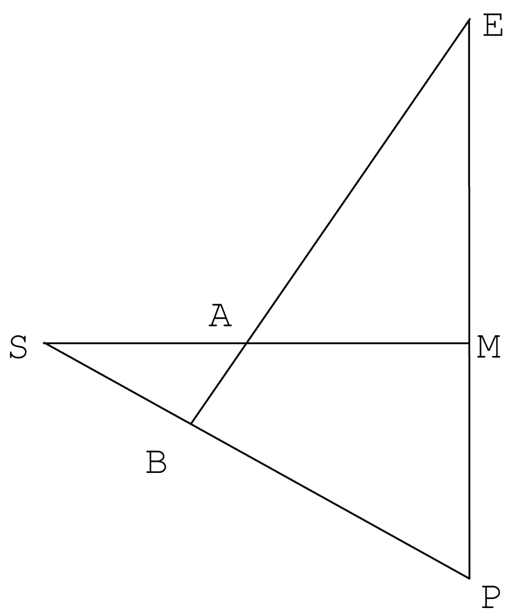


Figure 24.2: Summer solstice, abstract version

In order to determine the length of the day, we will find the arc AB of the equator. The reason that we do this is that the celestial equator is like a clock. This works in the following way. There are 24 hours in each day, and there are 360° in a circle. Observe that the quotient 360/24 is 15. This means that 15° of the equator rise each hour. Suppose, for example, that we find that the arc AB is 15°. That means that there is one hour of sunlight before B makes it to the horizon. There are then six hours for B to move from the horizon

Convince yourself that arc AE is 90°.

to the meridian at solar noon. Thus, the morning, sunrise to noon, would be seven hours. The whole day, then, would be 14 hours.

We must use specific information, the observer's latitude, as well as general information, the angle of the ecliptic to the equator. Those pieces of information are included in Figure 24.3.

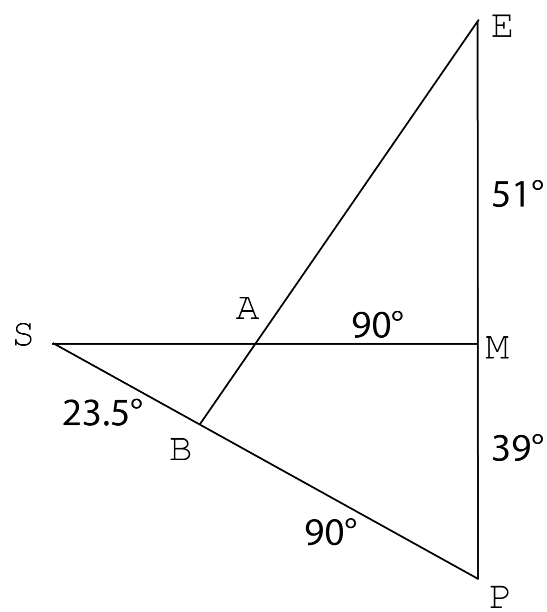


Figure 24.3: Summer solstice, labeled arcs

Since this is our first application of Menelaus's Theorem, there is one more figure to make everything clear, Figure 24.4. This is like the previous figure, but the lines have been reflected across a horizontal line. That makes the figure look exactly like the diagram for Menelaus's Theorem.

The arcs PM and EM are known, because the observer's latitude is known. The arcs AE and BP are each 90°, since they are arcs involving the equator, the meridian, a pole, and the horizon. The arc BS is about 23.5°. This is what is called the obliquity of the ecliptic. It is the same for all observers throughout the earth, and remains fairly constant over long periods of time.

Stand and point your left arm due east. Point your right arm straight overhead. Keep your left arm in place and slowly move your right arm towards the point on the horizon that is due south. Your arms make a right angle the whole time. That is an explanation of why AE is 90°.

Observe that we want to use the form of Menelaus's Theorem that tells us about ratios of parts with parts, since AB is the unknown thing that we seek, AE is a thing that we know, and each is a part of the arc BE. This means that we will use MT II. First, we will write down the ratios of chords using the letters.

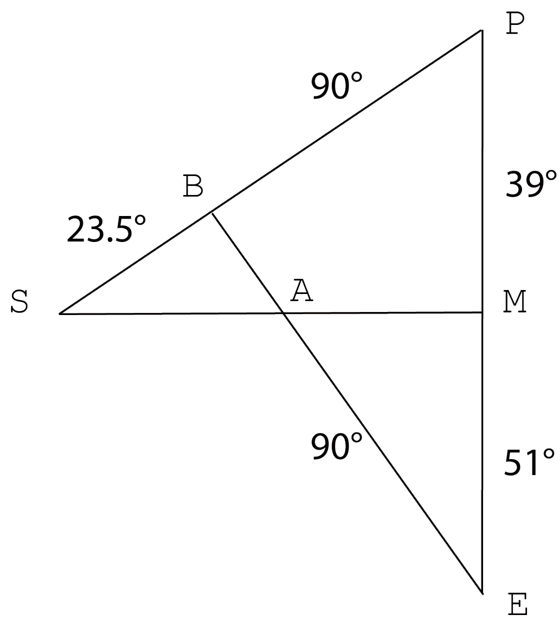


Figure 24.4: Summer solstice, reflected abstract figure

$$\text{crd } 2 \text{ arc } PM : \text{crd } 2 \text{ arc } ME$$

is the same as

$$(\text{crd } 2 \text{ arc } AB : \text{crd } 2 \text{ arc } AE).(\text{crd } 2 \text{ arc } PS : \text{crd } 2 \text{ arc } BS)$$

That is simply MT II, restated using different names for points. Now let us rewrite this using all of the information that we know.

In order to avoid cluttered notation, we will limit the use of parentheses. This means we must clarify an ambiguity. An expression like crd 2×39, which involves both the "chord" operation and the multiplication operation, should be understood as referring to the chord of the product. It should not be understood as 39 times the chord of 2 degrees.

$$\text{crd } 2 \times 39 : \text{crd } 2 \times 51$$

is the same as

$$(\text{crd } 2 \times \text{sought} : \text{crd } 2 \times 90).(\text{crd } 2 \times 114 : \text{crd } 2 \times 23.5)$$

Let us collect the required chord values in a table.

78, for example, comes from 2×39.

angle	chord
78	75.5
102	93
180	120
228	109
47	48

The computation of the chord of 228° requires a remark. For angles greater than 180°, we did not continue our chord table. Nonetheless we can still make sense of the chord of such an angle. If we call the large angle α, then the chord of α is the same as the chord of $360 - \alpha$, since the chord just depends on the endpoints, not which part of the circumference you call the arc. Therefore, we can compute the chord of 228° by instead computing the chord of 132°.

We then have this equality of ratios.

75.5 : 93
is the same as
(crd 2 × sought : 120).(106 : 48)

You can write this as an equation involving fractions, treating crd 2 × sought as an unknown. The conclusion is that crd 2 × sought is about 44. Now we will use our chord table in the other direction. We know the chord, and we want to look up the corresponding angle. We see that the angle 42° has a chord that is a bit too short, but the angle 45° has a chord that is a bit too long. Let's settle on 43°. Then twice the angle we seek is 43°, thus the angle we seek is 21.5°.

Now we need to interpret the angle in order to obtain a sense of the length of the day. The arc 21.5° on the equator corresponds to $(24/360) \times 21.5$ hours, in other words about 1.4 hours. Thus, the sun is above the horizon for 1.4 hours before the point on the equator that culminates (reaches the meridian) with the sun rises. Therefore, the time from sunrise to noon is 7.4 hours. By symmetry, the time from noon to sunset is also 7.4 hours. Thus, the length of the day is 14.8 hours. We usually use minutes and not fractional hours, so we can rewrite that by computing 0.8×60, which is 48. Thus, the length fo the summer solstice, at the latitude 39° is about 14 hours and 48 minutes long.

In the figures and calculations above, we used the name B for the point on the celestial equator that culminates—reaches the meridian—with the sun.

Our computation gives a result that is quite close to the reported length of the solstice at that latitude. It differs by a few minutes, however. What makes our computation inexact? First, it should be clear from our production of the chord table that many of the quantities we use are approximations rather than exact values. A second factor is that we computed using an abstract point in the sky, but the sun has a disc that is visibly extended. Sunrise and sunset times are reported not for when the center of the sun, the abstract point, rises, but instead for the times that the sun's disc first becomes visible and then finally disappears completely again. A third item to consider is that the earth's atmosphere acts like a lens, affecting the light that passes through it. As a result, our vision of the sun depends not only on its location within the abstraction that is the celestial sphere, but also on particular atmospheric elements that vary over time. All three of these factors lead to corrections that are small, however. The computations we complete using the method illustrated above give good results.

24.1.2 *Place of Sunrise*

We can use Menelaus's Theorem to reason about the sun in a different way. Instead of asking how long the day will be, we can inquire about where on the horizon the sun will rise. On the equinoxes the sun rises exactly on the equator, and so lies due east. In the summer and winter, though, the sun rises either to the north or to the south of due east. It is possible to determine the extent of the deviation from due east.

Let us again consider the latitude 39°, and let us now consider the winter solstice rather than the summer solstice. Where on the horizon will the sun rise on this, the shortest day of the year? We know that it will be in the southeast. How much south of due east will it be?

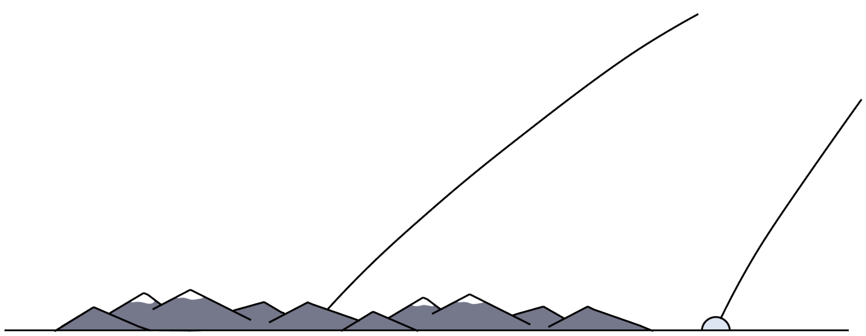


Figure 24.5: Winter Solstice

Consider Figure 24.5. This shows the sun rising on the winter solstice, with the celestial equator (the arc on the left) and the ecliptic (the arc on the right, through the sun) also depicted. A more abstract, labeled version of the diagram is in Figure 24.6. Once again, P is the south celestial pole, invisible to the observer below the horizon, and E is the intersection of the ecliptic and the equator. We are interested in the arc AS, which represents the extent to which the sunrise is south of due east.

For this computation we will use the first form of Menelaus's Theorem, MT I.

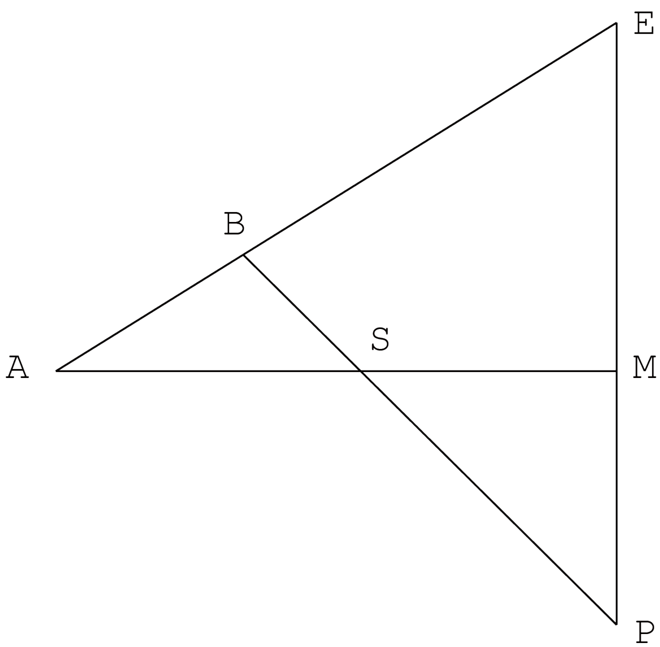


Figure 24.6: Winter Solstice, abstract version

crd 2 arc PE : crd 2 arc ME

is the same as

(crd 2 arc PB : crd 2 arc BS).(crd 2 arc AS : crd 2 arc AM)

That is simply MT I, restated using different names for points. Now let us rewrite this using all of the information that we know.

crd 2×90 : crd 2×51

is the same as

(crd 2×90 : crd 2×23.5).(crd $2 \times$ sought : crd 2×90)

Let us collect the required chord values in a table.

angle	chord
47	48
102	93
180	120

120 : 93
is the same as
(120 : 48).((crd 2 × sought) : 120)

Thus, the chord of twice the angle sought is $120 \times 48/93$, which is about 61.9. Considering the chord table, we see that the angle of 60° is too small, and 63° is slightly large. We conclude that the angle 62° has a chord of roughly the right size, so that the arc AS which we sought is about 31°.

24.2 Meridians, Solar Noon, and Terrestrial Latitude

We can get a reasonable sense of where east, west, north, and south lie by simply watching the sun throughout the year. The result that we obtain will only be approximate, however. This section explains how we can obtain a more exact determination of a meridian line and terrestrial latitude.

Here is a procedure to determine at any time of year a meridian line that is, in principle, exact. You could understand a proof of the method, but it would be somewhat difficult. It involves cones. This method is due to someone named Diodorus.

If you do not intend to carry out the procedure, you do not need to read through it closely. In any case, be sure to continue with the next section on solar noon.

Task: Produce, at a given location, a meridian line, i.e., a line that corresponds to the great circle through the celestial poles and the zenith.

Procedure:

1. Wait for a day when it will be sunny at least for a few hours.
2. Put a reasonably tall stick straight into the ground. Be sure that you have access to the space north of the pole—don't put it right next to a fence. Seek a location where the ground is relatively level.
3. In mid-morning, mark the end of the stick's shadow.
4. Around midday (it doesn't need to be exactly noon—you are not yet able to determine that), mark the end of the stick's shadow.
5. Sometime in the afternoon, mark the end of the stick's shadow again.
6. Consider the length of the first shadow and the third shadow. If they are exactly the same, you are done. The line between the two

points determined by those shadows runs exactly east and west. Produce the line through your stick which is perpendicular to this line. That is your meridian line.

7. Supposing that the first and third shadows are not exactly the same length, pick the smaller of the two, and call its length ℓ.

8. Find the hypotenuse of the right triangle whose legs are ℓ and the stick height. In other words, compute the square root of the sum of ℓ^2 and the square of the stick height. Call this r.

9. Refer to Figure 24.7. The left part refers to the shortest shadow, the one near midday. The right part refers to the longest shadow.

10. In each case, let the length AC be the measured shadow length, and let D be the point such that BD has the length r.

11. In each case, use similar triangles, the Pythagorean theorem, addition, and subtraction to compute the length CE.

12. Produce a line on the ground connecting the end of the shortest shadow and the end of the longest shadow.

13. Extend the shortest shadow by CE (as computed above) and shorten the longest shadow by the other computed value CE. Note that CE refers to something different in each case.

14. Produce the line connecting the extension of the shorter shadow and the shortened longest shadow.

15. Mark the intersection of the two lines you have produced.

16. The line between the end of the mid-length shadow (the one with length ℓ) and the point of intersection just constructed runs exactly east and west. Call it L.

17. Produce the line through the base of the stick in the ground which is perpendicular to the line L. This line is the meridian line.

24.2.1 Solar Noon

When you have a meridian line, you can determine solar noon exactly, provided that the sun is visible. Place a stick vertically in the ground on the meridian line. When the shadow of the stick lies along the meridian line, this is solar noon.

For many in the northern hemisphere, the shadow of the stick will always be to the north of the stick. Near the equator, though, the shadow might point to the south.

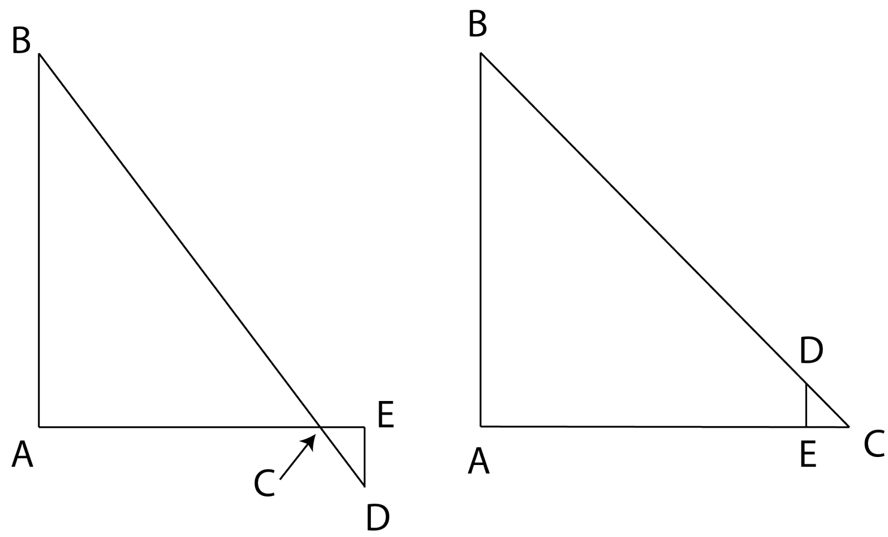


Figure 24.7: Computing shadow lengths

At solar noon on one of the equinoxes, you can use a meridian line to determine your latitude accurately. Simply determine the angles in the triangle made by the stick and its shadow on the ground. The bottom of the stick is at a right angle with the ground. The angle formed at the top of the stick, using the stick and the line from its end to the end of its shadow, is your terrestrial latitude. Note that this angle is the same as the angle between the sun and the point directly overhead.

24.2.2 *Size of the Earth*

Suppose that you determine the latitude of two locations, and that one of the locations is due north of the other. If you measure the distance between the two locations, you can determine the size of the earth, taking the earth to be shaped as a sphere. This is the basic idea of a calculation first done by Eratosthenes, who also developed the sieve for prime numbers you studied earlier. You might be able to compute in this way yourself when you travel to the north or south. For now, we simply record the result. The diameter of the earth is about 7925 miles, or 12750 kilometers.

These are approximations, not exact. The earth is only approximately a sphere.

24.3 *Near the Equator*

At latitudes relatively far from the terrestrial equator, such as those throughout the continental United States, the sun never appears at the zenith, directly overhead. Instead, at noon, when the sun reaches its highest point for the day, it remains somewhat to the south.

There is a special latitude called the Tropic of Cancer. This latitude, just south of the southernmost parts of the state of Florida, is about 23.5°, the same as the obliquity of the ecliptic. At such a latitude, the sun is directly overhead exactly on the summer solstice, and on no other day.

Consider the terrestrial equator in comparison. There, the celestial poles are exactly on the horizon, to the north and to the south, and the celestial equator passes directly overhead. Here the sun passes directly overhead twice each year. One time is on the vernal equinox, and the other on the autumnal equinox.

There is an analogous latitude in the southern hemisphere. It is called the Tropic of Capricorn.

At latitudes between the Tropic of Cancer and the equator, the sun passes directly overhead two times each year. Those days are near the solstice for latitudes near the Tropic of Cancer, and near the equinoxes for latitudes near the equator.

24.4 Near the Pole

The sun does not always rise exactly in the east. Instead, the place of rising depends on the latitude of the observer's location along with the time of year. Some unusual conditions arise when we consider locations that are near the terrestrial poles.

Consider the latitude 70°, and think about the location of the sun at noon on the winter solstice. At this latitude the pole is just 20° away from directly overhead. The celestial equator thus rises only 20° above the horizon in the due south. On the solstice, though, the sun is 23.5° south of the celestial equator, at the solstitial point on the ecliptic. This means that the sun is more than 3° below the horizon at noon. In other words, the sun does not rise above the horizon at that location on that day.

At another extreme, consider that same latitude at midnight on the summer solstice. The sun is 23.5° above the celestial equator, which is only 20° below the horizon in the due north, which means that the sun does not set on this day.

24.5 Exercises

Exercise 1

Determine the length of the summer solstice day at the latitude 45° N.

Exercise 2

Determine the length of the summer solstice day at the latitude 54° N.

Exercise 3

Determine the length of the summer solstice day at the latitude 60° N.

Exercise 4

Draw diagrams like those in Figures 24.1, 24.2, 24.3, and 24.4 for the summer solstice day length at the latitude 39° S, i.e., in the Southern Hemisphere. Which pole is below the horizon for this observer? Can you use earlier computations to determine this day length?

Exercise 5

Restate both parts of the Theorem of Menelaus (Theorem 113) with differently named points.

Exercise 6

At what northern latitude is the sunrise on the winter solstice exactly southeast (i.e., the angle is 45° south of due east)?

Exercise 7

Give a description, without using specific numbers or calculations, of the sun and its appearance directly overhead for observers at the latitudes 10° N and 20° N. Approximately when does the sun appear overhead for each one, and how many times, if at all?

Exercise 8

Consider an observer at the latitude 70° N. Discuss the behavior of the sun on the summer solstice. Where in the sky is it at various times of day? What happens at "night?"

Exercise 9

Suppose that you stand at the north (terrestrial) pole on June 21. What will the sun do? Be as specific as you can be. How will your shadow behave?

Exercise 10

What is the southernmost (i.e., least) northern latitude at which the sun does not set on the summer solstice? Do any people live that far north? Look at a map.

Exercise 11

Use a map to learn about the latitudes of Iceland. Describe what it is like to be in Iceland in June and in December.

25
A Refined Solar Model

So far we have used a simple model, the sphere, to describe the motion of the sun throughout the year. The ecliptic, a great circle on this sphere, is the path through which the sun passes over the course of a year. A first approximation, suitable for many purposes, is to think of the sun passing uniformly along this circle over the course of a year. Thought of in this way, the sun moves roughly one degree per day along the ecliptic. (There are 360 degrees in the circle, and about 365 days in a year.) We will now refine this, as the sun's motion is not quite uniform.

25.1 Differences among Seasons

The solstices and the equinoxes are distributed along the ecliptic at $90°$ increments. It turns out that the sun takes slightly different amounts of time to pass from one of them to another. This can be seen in the table below, which represents typical dates for each of the events (the dates can vary slightly from year to year).

The seasons here refer to the northern hemisphere.

spring equinox	March 20
summer solstice	June 20
fall equinox	September 22
winter solstice	December 21

These dates are all about three months apart. They are not, however, exactly the same amount of time apart.

Both March and May have 31 days, so the time from the spring equinox to the summer solstice in the table above is 92 days. July and August both have 31 days, so the time from the summer solstice to the fall equinox is 94 days (two months of 31 days, one month of 30 days, and two extra days). Continuing in this way, we find that the time from the fall equinox to the winter solstice is about 90 days, and the time from the winter solstice to the spring equinox is about 89 days.

A more precise accounting of the seasons is this one, which can be found by more precise measurements of solstice and equinox times, or by averaging over a number of years.

season	length
spring	92.75 days
summer	93.65 days
fall	89.85 days
winter	89 days

In order to account for this variation, we need to have a more sophisticated model, something that will provide for the observed non-uniformity of the seasons.

25.2 *Epicyclic Model*

We understand how to work with circles and chords of circles, so we will continue to use these tools. We will use them in a more complicated manner though. Now we will combine two circles.

We treat the observer as the center of a large circle. On that circle we place a little circle, called an *epicycle*. This little circle will rotate at the same angular speed, but in the opposite direction, as the larger circle. The motion you should envision is shown in Figure 25.1. The result of the epicycle's rotation is that a point on the epicycle stays in the same position relative to the center of the epicycle.

It is typical to call the large circle, the one on which the epicycle moves, the *deferent*. We will not use this term, but you might see it elsewhere.

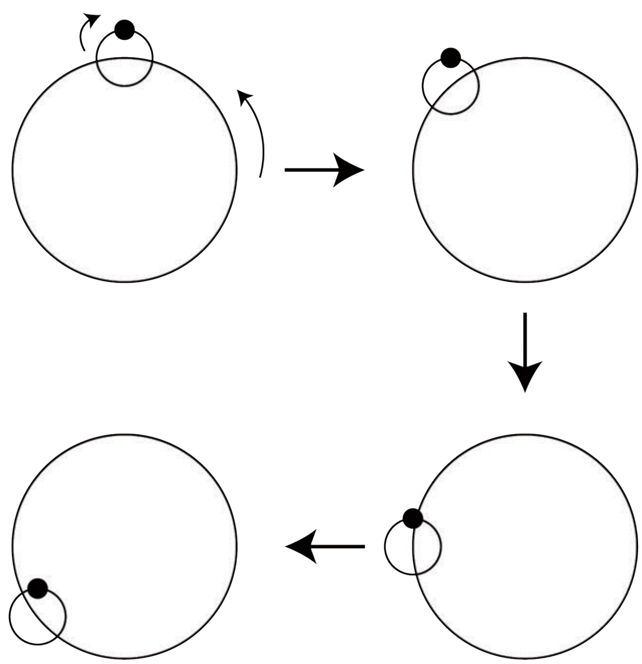


Figure 25.1: Example of epicycle

Let us observe how such a variation can account for seasons of different length. We think of ourselves as looking down upon the earth and sun from near the north celestial pole. The daily motion of the celestial sphere from east to west is clockwise in Figure 25.2. Each day, that circle (representing the whole celestial sphere) rotates clockwise, one time, about the dot at the center, which represents the earth. Over time, the sun falls back relative to the fixed stars, and therefore the motion of the sun within the celestial sphere is counterclockwise (seen from the north). In what follows, we will ignore the daily rotation of the fixed stars, and only think about the movement of the sun within the fixed stars over the course of a year.

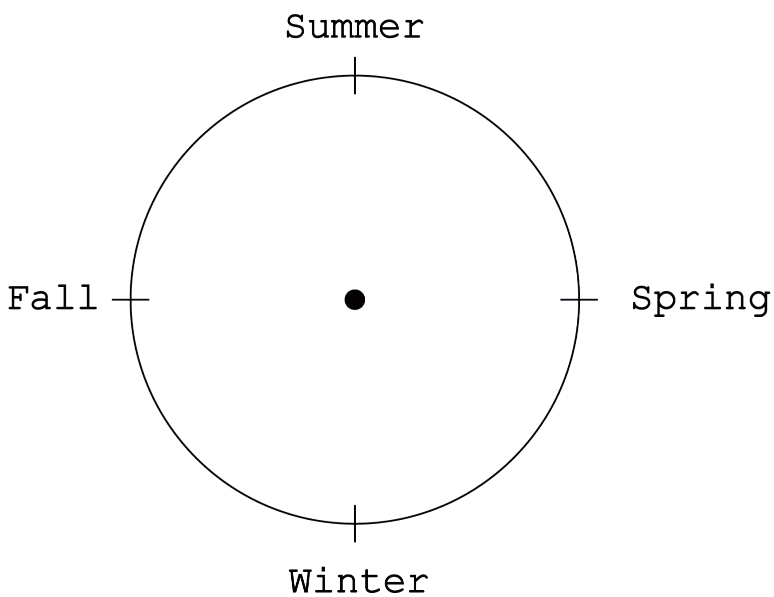


Figure 25.2: The ecliptic, seen from the north celestial pole

In order to use an epicycle model to obtain uneven season lengths, we need to determine two different parameters. One of them is the angle between a radius of the epicycle and the perpendicular lines determined by the equinoxes and solstices. This angle determines which seasons are short and which are long. The other parameter is the ratio of the radius of the epicycle to the radius of the circle that carries the epicycle. This ratio determines the extent to which the seasons vary in length.

We presume that the epicycle moves on the larger circle with a constant velocity. When the sun, carried on the epicycle, is ahead of the epicycle's center ("ahead" is said with respect to the counterclockwise motion of the epicycle), it enters a given quadrant early, meaning that it can spend more time there, since time is determined according to the uniform motion of the epicycle's center. When the sun is behind the center, on the other hand, it enters a given quadrant later. By reasoning in this way, and using our knowledge of the lengths of the various seasons, we can determine that the relationship between the epicycle model and the solstitial and equinoctial points of the ecliptic must look like Figure 25.3.

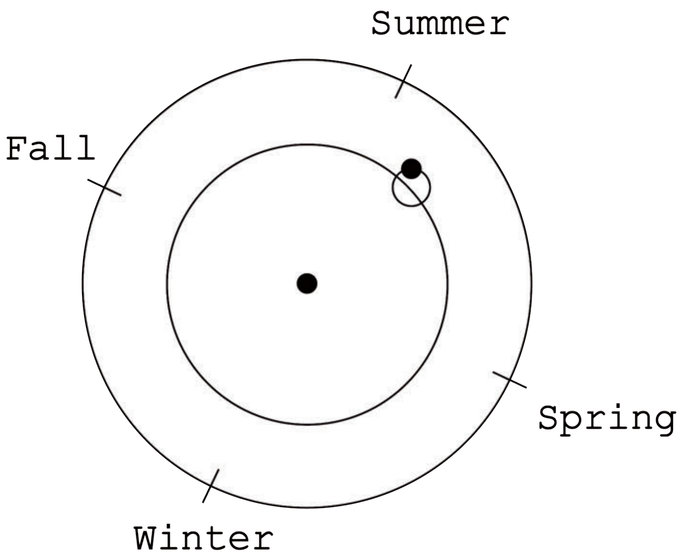


Figure 25.3: Relation of epicycle to solstices and equinoxes

In that figure you see that the sun is ahead of the epicycle's center when entering spring, and only slightly ahead of it when entering summer. Thus, the spring will be relatively long, as desired. Similarly, the sun is ahead of the epicycle's center when entering summer, and behind the epicycle's center when entering fall, so that the summer will also be long, as we have observed. We will check the fall and winter later.

We will now use our observed season lengths to compute the relative size of the epicycle. In order to convert days to angles, use the fact that the sun moves through 360° in roughly 365.25 days. This means that spring is about 91.4° and summer is about 92.3°, where the angle is the amount by which the center of the epicycle is displaced.

Proposition 114. *The ratio of the radius of the sun's epicycle to the radius of the circle that carries the epicycle is approximately* $1 : 29$.

Proof. Consider Figure 25.4. The point B is the sun at the spring equinox, and the point A is the location of the epicycle's center at the spring equinox (so AB is the epicycle radius). Similarly, the point D is the sun at the summer solstice, and the point C is the epicycle's center at the summer solstice. The points F and E are likewise the epicycle's center and the sun at the fall equinox. The angles BOD and DOF are both right, since they are the angles determined by the ecliptic arcs which correspond to spring and summer, and the solstices and equinoxes are evenly distributed along the ecliptic. The angle AOC, determining the length of spring, is about 91.4°, corresponding to the 92.75 days given earlier. The angle COE, giving the length of summer, is about 92.3°.

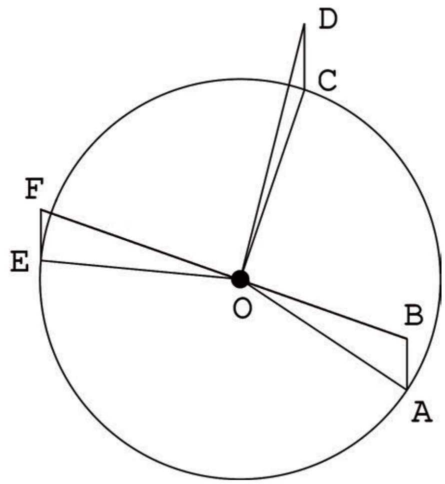


Figure 25.4: The sun's epicycle with labeled points

The angle corresponding to the arc ACE is then the sum $91.4 + 92.3$, which is 183.7. This means that the two angles AOB and EOF are, together, about 3.7°. It is possible to show that the angles AOB and EOF are the same. They are each, therefore, about 1.85°.

An exercise guides you through the proof that the angles are the same. For now, observe that BOF is a line (i.e., two right angles). Furthermore, since AB and EF are parallel, we can conclude using alternate interior angles that EFO and ABO are, together, two right angles. That is the beginning of the proof.

Consider Figure 25.5, with E and F indicating the same points as before. The point G is on OF, and so both the angles EGF and HGF are right. The chord EH corresponds to an angle that is twice the angle EOF. In other words, EH is the chord of 3.7°. The chord of 3.7° is roughly 4. This means that the segment EG is two units in length, when OE (a radius) is 60 units in length.

Now consider Figure 25.6. The point G is on the line OF as before, the point K is on the line OD so that CK is perpendicular to OD, and the lines OD and OF are perpendicular. Furthermore, the segments CD and FE (radii of the epicycle) are congruent. As a result, the triangles EGF and DKC are congruent. We want to determine the size of GF and CK.

The lines OD and OF are perpendicular because they come from a solstice and an equinox, and these are 90° apart on the ecliptic.

Since angle DOE is 92.3° and angle EOF is 1.85°, we conclude that angle COD is about 0.45°. Reasoning as before, we see that twice the angle COD has a chord which is twice CK. Thus, we are looking for the chord of the angle 0.9°, which is roughly 1. Half this is the segment CK. We then conclude that CK and GF are each roughly 0.5.

Finally, we use the Pythagorean theorem to determine the relative size of the epicycle. We approximate the square root of $2^2 + 0.5^2$ and obtain 2.06. Since the radius OE of the circle carrying the epicycle was taken to be 60 units, we obtain the ratio $2.06 : 60$, which is

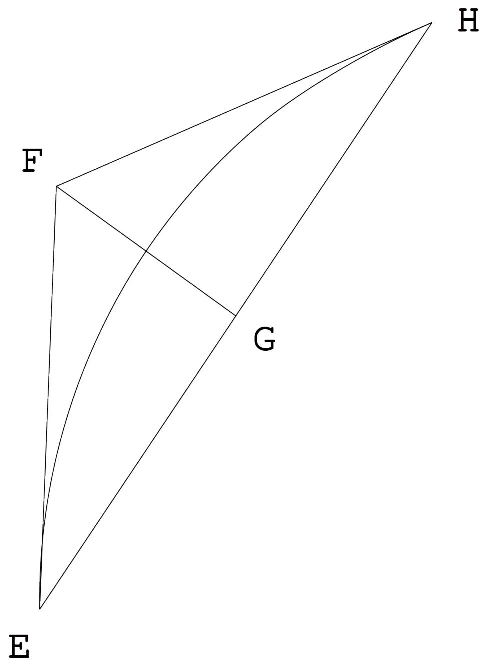


Figure 25.5: Finding the relative size of the epicycle

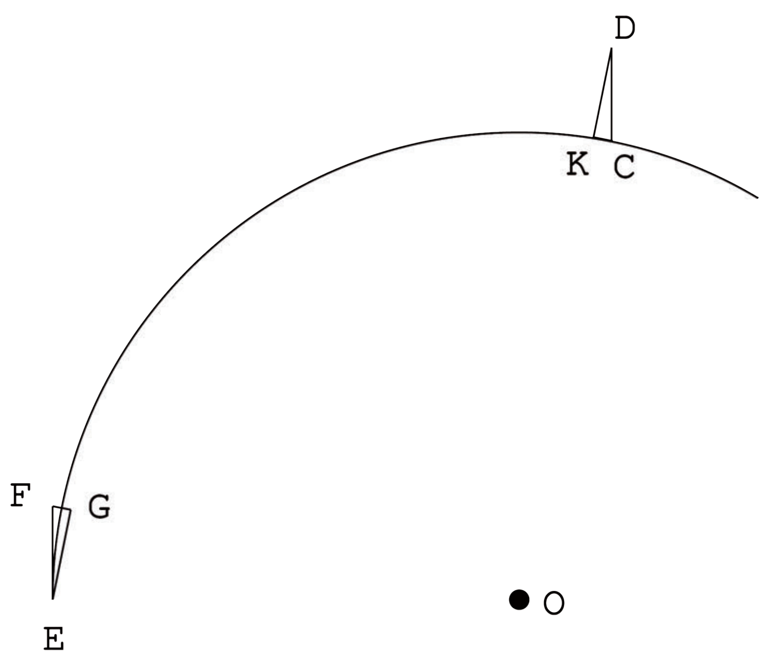


Figure 25.6: Configurations in summer and fall

roughly 1 : 29.

□

Given the relative sizes of the segments in the triangle EGF, we can compute the angles as well. Consider Figure 25.7. We see that twice the angle EFG has a chord which is twice the segment EG. We must rescale, though, to obtain suitable units, since our chord table presumes a diameter of 120 parts.

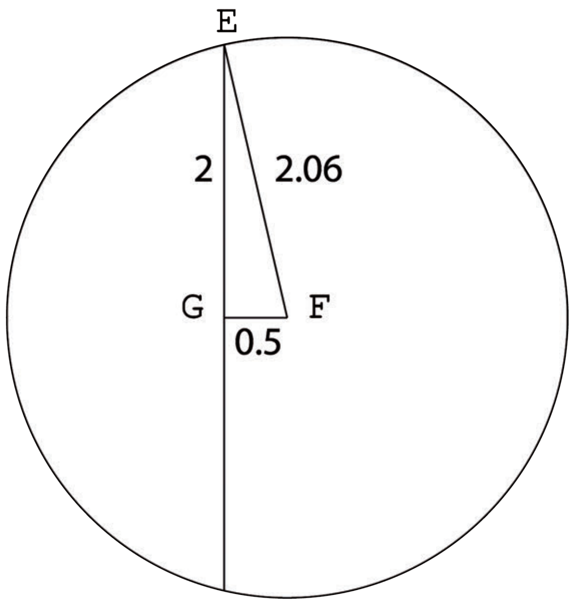


Figure 25.7: Computing the angle, solar epicycle

Divide 120 by 4.12 (twice the radius) to obtain the rescaling factor 29.12. The chord EG is 2 parts in the original units, so its double is 4 parts in the original units. Multiply 4 by 29.12 to obtain approximately 116. This is the chord of twice the angle EFG. We find in the chord table (Figure 23.11) that the angle 150° has a chord of about 116. Thus, angle EFG is about 75°.

This rescaling factor is the same as the ratio of the larger circle to the epicycle.

25.2.1 *Checking Fall and Winter*

Now we can use that angle between the fixed epicycle radius and the solstitial and equinoctial lines to determine what our epicyclic model says about fall and winter. This is a way to check whether what we are doing is sensible. If there is a significant disparity between what we observe and the numbers coming from the model, it suggests we should choose a different model.

Consider Figure 25.8. The points are labeled consistently with Figure 25.4, so that F is the sun at the fall equinox, for example. This figure is more to scale, so the epicycle radius is small. The point K is the epicycle's center at the winter solstice, and the point L is the sun at the winter solstice. The triangle LJK is right, with the right angle at J. The point J is useful for naming angles. The point H is introduced simply to illustrate the extent to which the epicycle radius deviates from the perpendicular lines of the solstices and equinoxes.

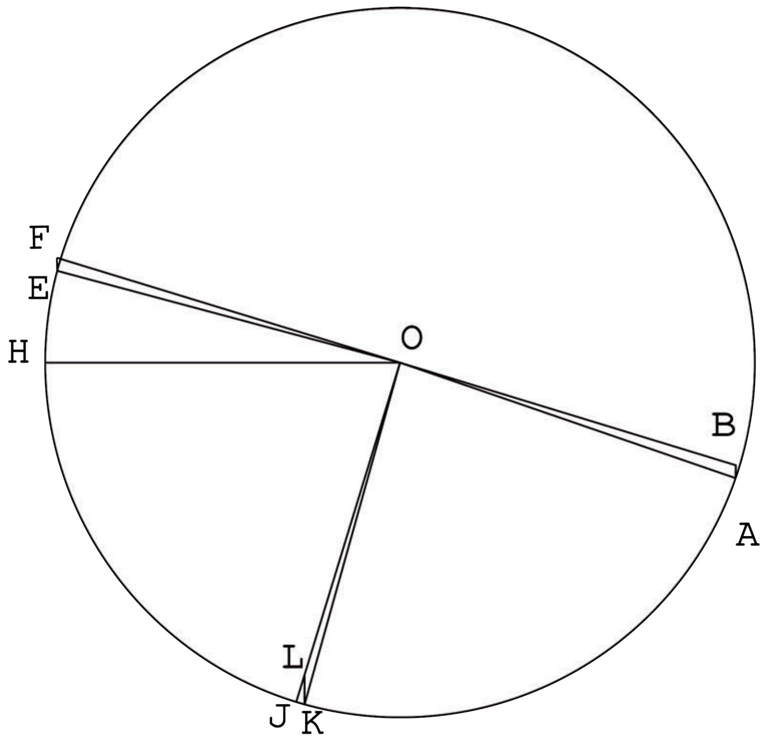


Figure 25.8: Fall and Winter, computation

Note that angle FOJ, the angle between the fall equinox and winter solstice, is 90°. The length of autumn is given by the angle EOK. We must find angle JOK in order to find EOK (the distance the epicycle travels between fall equinox and winter solstice). Consider the right triangle LJK with KL the radius of the epicycle. The shorter leg of this triangle, JK is about 0.5 in units in which OH is 60, as we found earlier. Thus, the chord of twice the angle JOK is about 1, and so the chord of JOK is about 0.5.

Combining the various pieces of information, we see that the angle EOK is about $90 - 1.85 + 0.5$, which is $88.65°$. We must convert the angle to days by using the fact that there are about 365.25 days for 360°. We obtain roughly 89.9, very close to our observed 89.85-day autumn.

We are taking angle FOJ, known to be right since it is the distance from an equinox to a solstice, subtracting angle FOE, which we computed earlier, and then adding angle JOK, which we computed now.

You can perform a similar computation to obtain 88.9 days for the winter (or you can hunt for a shortcut—there are only four seasons). Again, the result will be close to the observed value.

25.3 Eccentric Model and Its Equivalence to Epicyclic

It is good to be able to think about a thing in more than one way. We can do that with the sun's non-uniform motion. Our second model will involve only a single circle rather than two.

25.3.1 The Model

This model of solar motion is called *eccentric* because it places the center of a circular motion away from the earth. The earth is "out of" (*ex*) the center. The sun is presumed to move with uniform speed around this circle centered on a point other than the earth. This is shown in Figure 25.9, where the earth is the small dot slightly away from the intersection of the diameters.

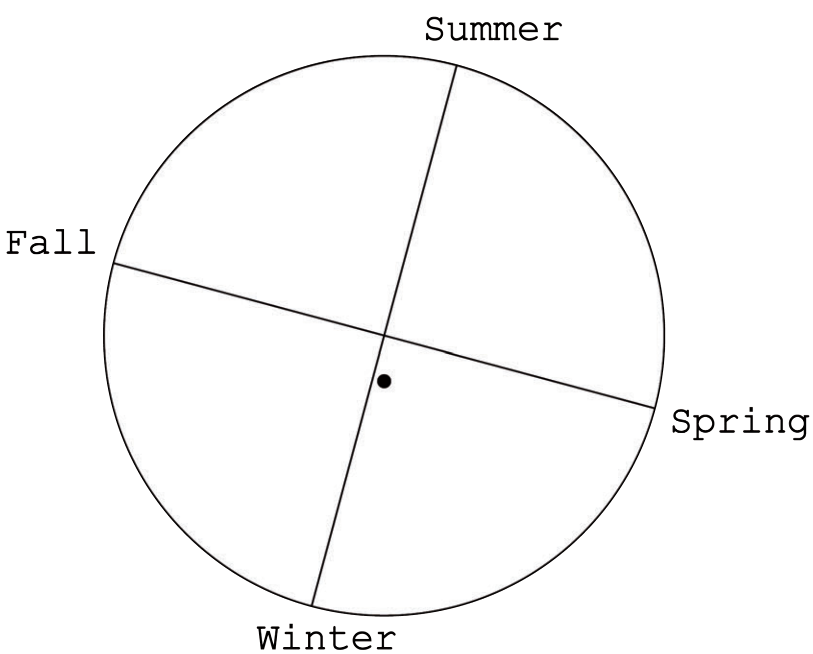


Figure 25.9: Eccentric model

Just as we did with the epicyclic model, it is possible to use the observed season lengths to establish relevant ratios and angles. For this model there are again two things to determine. The solstices and equinoxes determine diameters of the circle. The angle made by the diameter through the earth with those other diameters is one thing to determine. The second thing to determine is the extent to which the earth is displaced from the center. In other words, we need to find the ratio of the segment between the earth and the center to a radius.

The methods for determining these values are similar to what we saw in the epicyclic model. In particular, the ratio of the earth's displacement to the circle's radius is about $1 : 29$, the same as the epicyclic radius ratio.

25.3.2 *Equivalence*

Given two models for the same phenomenon, we can ask whether they agree in all cases or only in some. If the models agree in all cases, we say that they are equivalent.

Proposition 115. *Consider an epicyclic model and an eccentric model, subject to the condition that the radius of the epicycle is the same as the eccentric model's central displacement. Then the two models describe the same behavior.*

Proof. Consider Figure 25.10. The point E represents the earth, and the point C is the center of the circle for the eccentric model. The segments AB and CE are equal. Segment AB is a radius of the epicycle, and CE is the displacement of the earth from the center of the eccentric model circle. The radii EA and CB are the same. The angles DCB and DEA are the same, for we are considering the solar motion based on the passage of time corresponding to that angle.

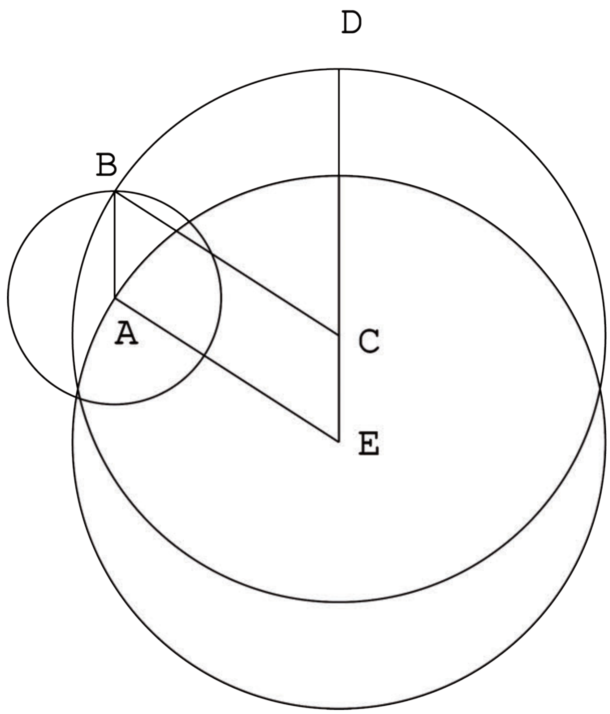


Figure 25.10: Equivalence of models

The models are equivalent because $ABCE$ is a closed figure. In other words, there is not a point B' distinct from B on the circle centered at C with radius CD in the form of Figure 25.11.

You will fill out this argument in greater detail in an exercise.

□

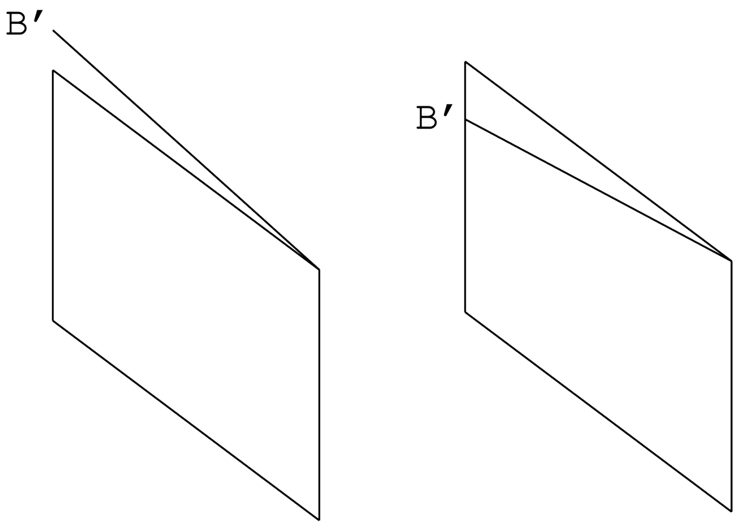


Figure 25.11: Impossible cases

25.4 Distances

Our models for the sun's motion suggest that the distance between the earth and sun is not constant but instead varies over time. This has been observed to be the case. Even though models more complicated than ours are now used to describe the relationship between the earth and sun, these newer models retain key features of our epicyclic model, including the variation in distance.

Refer to Figure 25.3. Observe that the sun will be furthest from the earth slightly after the summer solstice. At that time the radius corresponding to the sun on the epicycle points directly away from the earth. Similarly, the sun will be nearest to the earth shortly after the winter solstice. Compare to Figure 25.9, depicting the eccentric model.

25.5 Historical Variations

The lengths of the seasons have been observed to vary over time. In Ptolemy's *Almagest*, written almost 2000 years ago, spring is reported to be longer than summer. This is different from what we have used here to describe the present day. Ptolemy reports a spring of 94.5 days and a summer of 92.5 days. His fall and winter are about 88 and 90 days, respectively.

At an intermediate time, around roughly 1200 AD, the exchange of season lengths took place. Before that time seasons were as Ptolemy observed, with spring the longest season, followed by summer, then winter, then fall. Briefly, at about 1200 AD, spring and summer were of the same length, and fall and winter were also of the same length. Since then, for the past 800 years or so, summer is longest, then spring, then fall, then winter. According to the models that are used now, spring will continue to shorten and fall will continue to lengthen, but it will be many hundreds of years before those seasons are of equal length.

25.6 *Exercises*

Exercise 1

Convert the lengths of spring (92.75 days) and summer (93.65 days) to degrees, using the fact that there are about 365.25 days in a year, and 360° in a circle.

Exercise 2

We found the angle between the epicycle radius and the solstitial and equinoctial diameters. More specifically, following Proposition 114, we showed that the angle called EFG is approximately 75°. Is this an underestimate, or an overestimate? Explain.

Exercise 3

The argument in the proof of Proposition 114 that the triangles EGF and DKC are equal was somewhat abbreviated. Elaborate on it. Those triangles are in Figure 25.6.

Exercise 4

Refer to Figure 25.8. Determine the angle AOK. Interpret this angle in terms of time, converting it to days. Check that you obtain the number of winter days given in the text.

Exercise 5

Use the epicyclic model to determine, at least approximately, dates when the sun is furthest from the earth and nearest to the earth.

Exercise 6

Use other resources to find dates called *perihelion* and *aphelion*. Compare these dates to what you found in the previous exercise.

Exercise 7

Use Ptolemy's season lengths to establish an epicyclic or eccentric model for the sun's motion. The radius of the epicycle (or displacement of the earth from the eccentric center) will have the ratio roughly $1 : 24$ to the larger circle's radius.

Exercise 8

Complete the proof of Proposition 115, that the epicyclic and eccentric models are equivalent, using the following geometrical reasoning. The situations being excluded are depicted in Figure 25.11

1. Consider a point B' on the circle centered at C with radius CD, such that angle $B'CD$ is the same as angle AED and segment $B'C$ is the same as AE. (At this point we do not assume B and B' are the same.)

2. Continue to assume (as was done in the proposition) that AB and CE are equal and parallel.

3. Prove that CB' and EA are parallel. (Use alternate interior angles.)

4. Using a property of parallelograms, show that AB' is parallel to and equal to CE.

5. Conclude that B' and B must name the same point.

Effectively, the point B is the sun's location in the epicyclic model, and the point B' is the sun's location according to the eccentric model. The equality of angles asserted here amounts to saying that we have allowed the same amount of time to pass in both models.

Exercise 9

Complete the proof of Proposition 114 by showing that angles AOB and EOF are the same. The angles ABO and EFO combine to make two right angles, using Proposition 21 (I.27) as noted above. Consider the circle with center E and radius EF. It intersects the segment OF in a point X. The triangle EFX is isosceles. Use this to show that angle EXO is the same as angle ABO. Argue that triangle OEX is congruent to triangle OAB. Conclude that the angles at O are the same.

Euclid's Proposition III.2 implies that a line intersects a circle in at most two points. That result can be used here.

26
Terms of Time

Our careful consideration of the sun's motion means that we can now investigate the meaning of simple words, discovering in them a subtlety that is perhaps surprising.

26.1 Days

What is a day? Or, how do we know that a day has passed? In one sense this is a very simple question. At all but the most extreme latitudes, we experience sunrise and sunset each day. Each time that we see another sunrise, we know that a new day has come.

The subtler question is how to know *exactly* one day, rather than *at least* one day. We will consider three different senses of the word "day."

Definition 116. *A real solar day is the time from solar noon on one day to solar noon on the next day.*

Solar noon is the time that the sun is due south, for observers at latitudes above roughly 23.5° in the Northern Hemisphere. More generally, it is the time that the sun passes the meridian, the great circle determined by the celestial poles and the zenith. At solar noon, shadows shift from pointing somewhat west to pointing somewhat east.

Definition 117. *A sidereal day is the time it takes for a specific star to return to the meridian.*

In this definition we refer to stars other than the sun.

The stars all stay in the same configurations relative to one another over long periods of time. Therefore, the definition of sidereal day is independent of the choice of a specific star.

Let's compare these two kinds of day. Remember that while the motion of both the sun and the stars is from east to west, the sun tends to fall back slightly with the passage of time. It falls back roughly one degree per day. That means that if a star crosses the meridian at solar noon on one day, it will make it back to the meridian before solar noon on the next day. Over the course of the whole year this adds up, and the celestial sphere makes one more revolution than the sun.

Sidereal days are presumed to be of the same length, by the uniformity of the motion of the stars each night. Real solar days, on the other hand, vary in length, for reasons we will investigate later.

The sidereal and real solar days come directly from observation. There is a third type of day that does not arise immediately from observation but instead depends on a model. This third notion of day is the mean solar day. Suppose that an epicyclic model for the sun is established. The center of the epicycle can be called the "mean sun." The sun is sometimes ahead of this point and sometimes behind it.

In fact "mean sun" is more complicated. We also need to take into account the obliquity of the ecliptic.

Definition 118. *A mean solar day is the time it takes for the mean sun to return the meridian.*

A real solar day is sometimes longer than a mean solar day, and sometimes shorter. Each mean solar day, on the other hand, is of the same length. Here is a second way to think about what a mean solar day is. After many years we can see that the number of days in a year is roughly 365.25. A mean solar day is the amount of time such that when it is repeated 365.25 times, it is the same as a single year.

26.2 Hours

We all understand that there are, by convention, 24 hours in each day. The previous section, though, has shown that the notion of "day" itself requires some careful thought. So too with hours.

One notion of "hour" that is not useful for us is this—let each day and night have 12 hours each. This is complicated for a number of reasons. First of all, days and nights are not ordinarily of equal length. Days are longer than nights in the summer, and in winter, nights are longer. Thus, an hour of the night and an hour of the day will be different, with this sense of hour.

A more complicated phenomenon is present even on the equinoxes, when day and night are roughly equal. The sun has a visibly extended disc, which means that the sun is visible before the center of its disc rises and remains visible even after the center of its disc sets. Thus, even at an equinox the days and nights are not of the same length.

Here are two reasonable notions of "hour."

The word "equinoctial" means "at an equinox," either in the spring or in the fall.

Definition 119. *An equinoctial hour is one twenty fourth part of an equinoctial solar day.*

A solar day can differ from a mean solar day, being either longer or greater, so the following definition is also sensible.

Definition 120. *A mean solar hour is one twenty fourth part of a mean solar day.*

When we do not qualify the word "hour" it will mean "mean solar hour." Similarly, if we speak of "minutes" and "seconds," these will refer to portions of mean solar hours (i.e., the 60th and 3600th parts, respectively).

26.3 Years

We observe the passage of years with the passage of seasons, the varied weather arising from the variation in the sun's noontime elevation. If we simply estimate, counting the number of days from when the sun is at its solstice to the next time it is, we get 365 days in a year.

It is challenging to determine the exact time of the solstice.

Instead of considering a single year, we can consider many years. By considering longer periods it is less important to get the time of the solstice exactly right. We find that there are about 1461 days in four years, or 7305 days in 20 years. We can then call the quotient, 365.25, the number of days in a year.

This number too, like 365, turns out to be somewhat inaccurate over long periods of time. In 400 years, for example, we find that there are about 146097 days rather than 146100.

Definition 121. *A solar year, also called a tropical year, is the period of time from one summer solstice to the next.*

The definition could also be made using the winter solstice, or one of the equinoxes.

With days, we distinguished between solar and sidereal days because the relation between the sun and the background stars changes over time by roughly one degree per day. This means that we could also offer a second definition of year, the sidereal year, according to the time taken for the sun to return to a fixed star. We have no reason now to distinguish between a solstice (a point on the ecliptic furthest from the celestial equator) and a point in the celestial sphere, a star. In the final chapter on astronomy we will return to this theme.

26.4 The Equation of Time

To what extent does a real solar day differ from a mean solar day? We will examine this question briefly, using our epicyclic model and our knowledge of chords. There are two factors that contribute to the difference. One is the eccentricity of the sun's motion, the fact that it is not uniform about the earth. The second is the obliquity of the ecliptic. The term "equation of time" refers to the relationship between time as measured by uniform clocks and time as measured by the visible motion of the sun.

26.4.1 Contribution of Eccentricity

We will look down at the earth and sun from near the north celestial pole, as we did before. The daily rotation of the celestial sphere is then clockwise, and the slower motion of the sun is counterclockwise relative to the sphere, when the sphere is taken to be fixed.

Proposition 122. *When the sun is at its greatest distance from the earth, the epicycle reduces the real solar day by about 8 seconds from the mean one.*

Proof. Consider the point when the sun is furthest from the earth, as well as the sun one day later. These are depicted in Figure 26.1. Suppose that solar noon on both days is depicted. Recall that mean solar time regards the location of the mean sun, the center of the epicycle (depicted with a small dot), while the real solar day depends on the observed location of the sun (depicted by the larger dot).

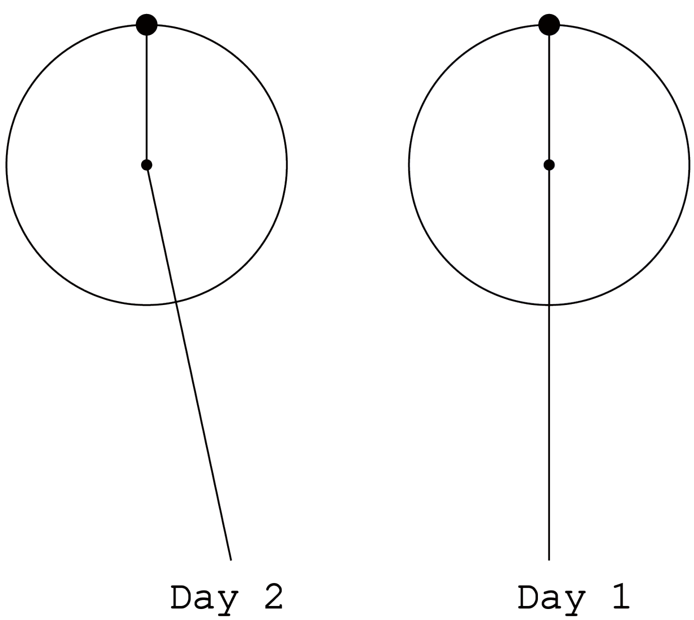


Figure 26.1: Sun at farthest point

On Day 1, the mean sun and the observed sun are at the meridian at the same time. On Day 2, on the other hand, the real sun leads the mean sun (where we consider the daily clockwise rotation of the whole celestial sphere). This means that the real solar day that passes from noon on Day 1 to noon on Day 2 is less than a mean solar day.

Day 2 is shown at a slight angle, and to the left, since we are considering both the configurations relative to a fixed celestial sphere, whose daily motion is clockwise. In one day, the sun falls back slightly relative to this sphere.

The real sun gets to the meridian faster than the mean sun. We can estimate the time difference.

In Figure 26.2 we see the sun on Day 2, when it is slightly ahead of the mean sun. The earth is at E, the sun is at S, and AS is a radius of the solar epicycle. The angle AEC is about 1°, since the mean sun falls back relative to the fixed stars by a degree each day (more precisely by the quotient $360/365.25$ of a degree each day). We want to determine the angle AES, which tells us the extent to which the real sun is ahead of the mean sun.

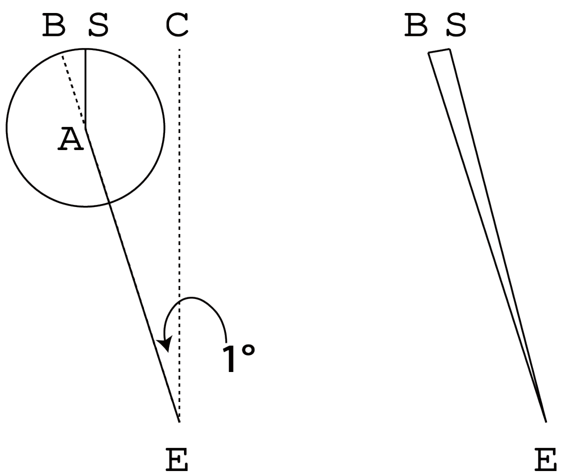


Figure 26.2: Real sun ahead of mean sun

Let EA be extended to the point B as shown. Then angle BAS is the same as angle AEC. First, take a system of units in which AS is 60 units, so that the epicycle has a diameter of 120 units. The segment BS is then roughly 1 unit. (Chords of small angles are roughly the same as the angles themselves.)

Now consider the triangle SBE. While angle SBE is not exactly right, it is near enough that we can treat it as such. We need a new system of units. Choose a unit so that AE is 58 units. Then AB is 2 units, by the ratio for the epicycle's radius. In this system of units, EB is 60 units. How big is BS? We have rescaled by a factor of 30, since we now take AB and AS to be 2 rather than 60. Thus, BS is, in this new system of units, $1/30$.

The ratio $2 : 58$ is the same as the ratio $1 : 29$.

Since small chords and angles are roughly the same, we conclude that angle BES is roughly $1/30^\circ$. Now we need to see what that means in time. The celestial sphere goes through about 360° in 24 hours, thus it goes through 15° in each hour. That means that there are 4 minutes for each degree. That means that the time we are considering, the time by which the real solar day is shorter than the mean solar day, is about one thirtieth of four minutes. Four minutes are 240 seconds, so the time difference is about 8 seconds. □

It is possible to reason like this about other times of the year. At other times the contribution tends to be smaller. When the real sun, the mean sun, and the earth form a right angle, the epicycle does not contribute meaningfully to variation between the real and mean solar days.

26.4.2 *Contribution of Obliquity*

The sun does not travel along the celestial equator. Instead, it travels on the ecliptic, and is sometimes to the north of the equator and sometimes to the south. The real sun moves uniformly along the ecliptic. The mean sun, on the other hand, is thought of as moving along the equator.

The mean sun gives the average motion. Since the sun is in the north for half the year and the south for the other half of the year, its average position is on the equator.

We will consider two kinds of variation due to the obliquity of the ecliptic. The first is when the ecliptic is at the largest angle with the equator, at an equinox. The second is when it is effectively parallel with the equator.

Proposition 123. *At an equinox, the obliquity of the ecliptic reduces the real solar day by almost 20 seconds relative to the mean solar day.*

Proof. Observe Figure 26.3, in which A designates the spring equinox. The arc AB is along the equator and the arc AC is along the ecliptic. Since we are not concerned here with exact values, and the arcs are small relative to the whole sphere, we can make a simplifying assumption. Rather than working on the sphere we can work on a plane. The chord of 1° is roughly 1, so we have a right triangle with hypotenuse of length 1. The chord of 23.5° is about 24.5. We will simplify and call it 25. Rescaling, since AC is 1 and not 60, we find that BC is about $25/60$, or $5/12$.

You can do Euclidean geometry in the sand at the beach even though the earth is spherical. The reasoning is the same here.

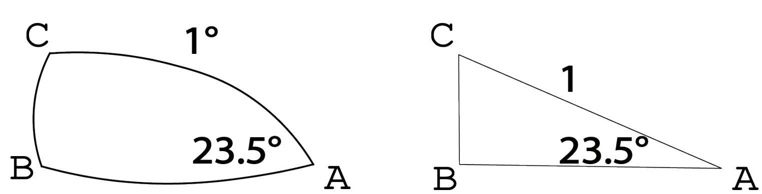


Figure 26.3: At the spring equinox: on the sphere, and a planar approximation

We will now show a way to simplify the estimate. You will proceed more exactly in an exercise.

The ratio $5/12$ is not much greater than $5/13$. The reason this is a useful observation is that $5^2 + 12^2$ is 13^2. Thus, we find that AB is roughly $12/13$. We avoided extracting a square root.

This means that when the sun travels one degree along the ecliptic, this amounts to only about $12/13$ degree along the equator. Thus, the

The real sun is moving from A to C. On the first day, the real and mean sun are both at taken to be at A. On the second day, the real sun is at C. The mean sun is one degree away from A, to the left in the diagram. We see that this is farther to the left than the point B.

real solar day will be shorter than the mean one by the time that corresponds to $1/13$ degree. Since one degree corresponds to 4 minutes, or 240 seconds, we find that this $1/13$ degree is about 18.5 seconds. □

The second variation to consider is near the solstice. At this time, the sun's daily regression relative to the fixed stars is effectively parallel to the celestial equator.

Proposition 124. *At the solstices, the obliquity of the ecliptic increases the length of the real solar day by about 20 seconds relative to the mean solar day.*

Proof. The solstitial points are the two points on the ecliptic at the greatest distance from the equator. Near such a point, motion along the ecliptic is almost exclusively in a plane parallel to the one defining the celestial equator. We will make a simplifying assumption and treat the motion as if it were exclusively in such a plane.

In Figure 26.4 the earth is at the center of the sphere, the arc through S and T is the ecliptic, and the arc through A and B is the equator. The solstice is at S, and the sun's location the next day is at T. The north celestial pole is at P, and the points A and B are on the great circles passing through P and the points S and T, respectively. Arc ST is one degree. The chord of arc ST is approximately 1, in units in which the sphere has a radius of 60 units.

Consider Figure 26.5, which shows a cross-section of the celestial sphere. The center of the sphere is O, the points S and A are as before, and T and B are not shown. We will find the segment OC.

To see why we want to find the segment OC, consider a different perspective on the celestial sphere, Figure 26.6. Here we look from above the north celestial pole. We assume that the circle on which S and T lie has been made by a plane perpendicular to the line formed by the north and south poles. The segment OC of the previous figure is now the radius PS.

You might find it helpful to think about this as if it were on the earth. The points S and T are at a higher latitude, while A and B are on the equator. The distance between two lines of longitude gets smaller as you get nearer the pole.

We return to the triangle of Figure 26.5. By similar reasoning to what we have done earlier, we find that OC is about 54.8 in units in which OS and OA are 60 units.

Return to Figure 26.6. Consider the triangles PST and PAB, which are similar. The chord ST is roughly 1, as we saw earlier, so by the ratio of the similar triangles we find that AB is roughly $60/54.8$, which is about 1.09. Since the angle APB is small, we can say it is roughly the same as its chord 1.09 as well. This means that the real sun falls back by 0.09 degree more than the mean sun. Converting 0.09 degree to time, we find an approximation of 0.09×240 seconds, which is 21.6 seconds. □

There are some important details that are glossed over here. We are finding a *difference* of angles. Recall the remarks of Section 23.3.3 to see the delicacy of this matter. It turns out that a better estimate of the chord of 1° is 1.04 rather than 1. We then would need to find the angle whose chord is $60/54.8 \times 1.04$. Then we would find the difference between that angle and 1°. Why is there no problem? Although we use fairly coarse approximations of the chords of angles that are close to one degree, the approximations are both coarse in the same way (both are underestimates by roughly the same amount) and the errors mostly cancel each other out.

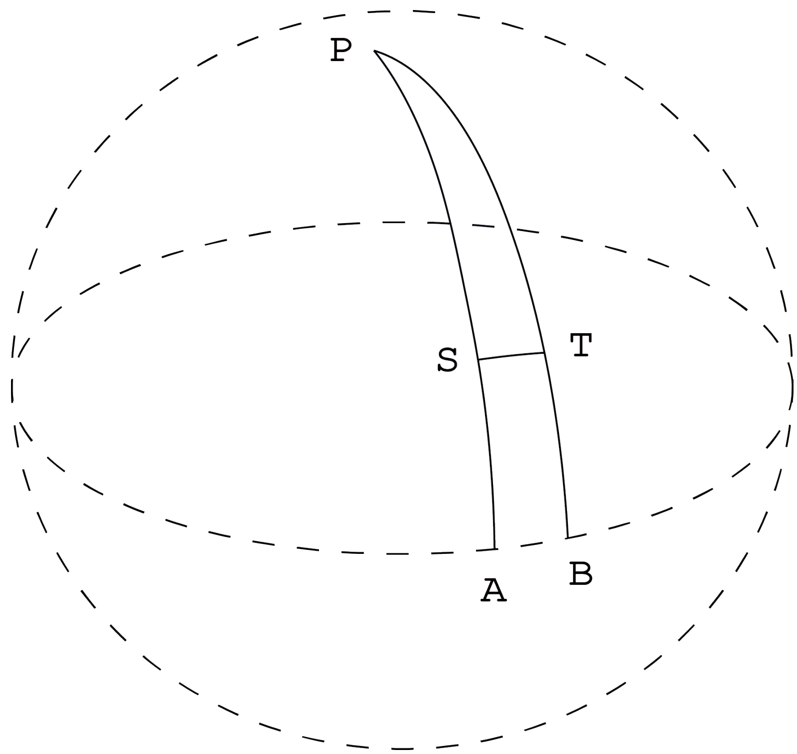


Figure 26.4: At the solstice

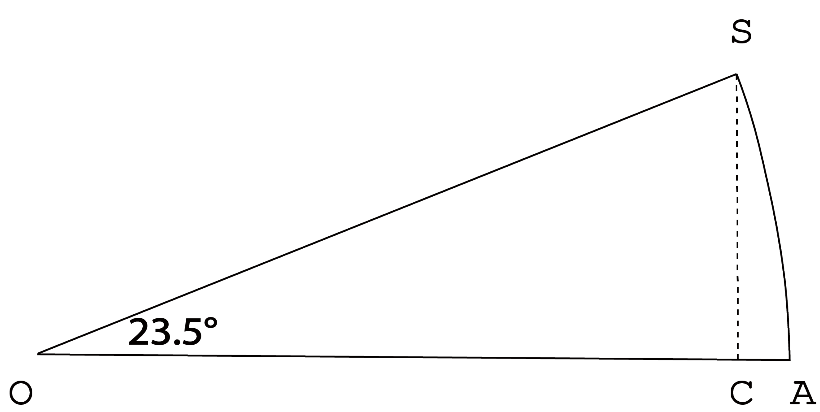


Figure 26.5: At the solstice, cross-section

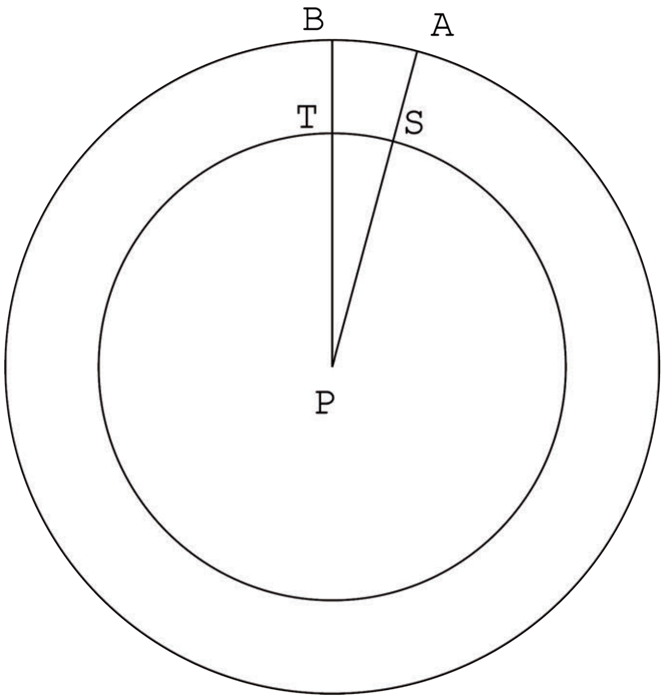


Figure 26.6: At the solstice, seen from above

26.5 Exercises

Exercise 1

Estimate the amount of time it takes for the disc of the sun to rise.

Exercise 2

One reasonable notion of "hour" that we did not define is a "sidereal hour." Give a definition of this term. Determine how it compares to the other hours we defined.

Exercise 3

Draw the sun against a background of stars at solar noon on consecutive days. You do not need to depict constellations exactly (though you should do that if you are able). Instead, the purpose is to show the change in position of the sun relative to the background stars. Show what it means for the sun to "fall back" relative to the stars.

The stars are not, of course, ordinarily visible at solar noon.

Exercise 4

Explain how the word "day" should be understood when we speak of the number of days in a year. It does not need to be one of our exact notions. It can involve counting sunrises.

Exercise 5

There are about 146097 days in 400 years. Based on this number of days, how many days would you say are in a year?

Exercise 6

In the Julian calendar, each fourth year was a leap year. The Gregorian reform of the Julian calendar made every fourth year a leap year, except for century years that are not divisible by 400. Explain this reform in light of the previous exercise.

Exercise 7

In the equation of time computation, Proposition 122, show that BS is roughly one unit when AS is 60 units.

Exercise 8

Using the epicyclic model, illustrate the two times that the sun, the mean sun, and the earth form a right angle. Show that the difference of day length due to the epicycle is negligibly small near such points. At what time of year does each occur? Be as specific as possible.

Exercise 9

The two solar models, the epicyclic and the eccentric, give the same description of the sun's position, but in different ways.

a.) Explain why it is more convenient to use the epicyclic model when talking about mean solar time.

b.) Explain what a mean solar day is, referring only to the eccentric model and not the epicyclic one.

Exercise 10

Compute the variation in the day length due to the epicycle when the sun is nearest to the earth. Compute the size of the arc, at least approximately, and convert it to time. Determine whether the real solar day is longer or shorter than the mean one.

Exercise 11

We used a simplifying approximation in the proof of Proposition 123, saying that the ratio $5 : 12$ and the ratio $5 : 13$ are roughly the same. Improve that computation by extracting a suitable square root. Determine how much that change affects the conclusion about day length.

Exercise 12

Draw a diagram like Figure 26.3 to illustrate how the obliquity of the ecliptic affects the real solar day at the fall equinox. What is different? (Hint: where will the next sunrise, a point on the ecliptic, occur relative to the arc marking the horizon?) Argue that the same reasoning that we used in the proof of Proposition 123 also applies here.

Exercise 13

There are 24 hours in a mean solar day, and roughly 365.25 mean solar days in a year. Determine the length, in mean solar hours, of a sidereal day. (There is a difference of one day between the number of solar days in a year and the number of sidereal days. Make sure that you know which one is the larger number. Then use a ratio.)

27
Elements of Lunar Astronomy

Having described key elements of the sun's motion, we will now study the moon. The appearance of the moon—whether it looks fully or partially illuminated—depends on its position relative to the sun.

27.1 Lunar Position by Phase and Season

The moon and sun both travel within the ecliptic, a great circle in the celestial sphere that is at an angle to the celestial equator. They sometimes appear higher in the sky, and sometimes lower, due to the obliquity of the ecliptic. Two of our astronomical principles are that the sun falls back relative to the fixed stars, and the moon falls back relative to the sun. These two principles let us explain where the moon will appear with a given phase in a given season.

These principles arose from your observations. Recall the list given in Section 22.2.

The moon is full when it is directly opposite the sun. When the moon is near the sun, we call it a new moon, and it is not visible, because it is hidden from our view in the sun's light. The new moon and full moon occur about two weeks apart. About one week after the new moon and one week before the full moon is the moon that we call a first quarter moon. At northerly latitudes, the moon and sun tend towards the south, and in this case the first quarter moon is illuminated on the right. It crosses the meridian, reaching its highest point, around sunset. We can be more specific about how the moon's first quarter location depends on the seasons, as illustrated by the following proposition.

Proposition 125. *At northern latitudes, the first quarter moon is high in the sky at sunset in the spring, and low in the sky at sunset in the fall.*

Proof. In the spring, the sun is near the spring equinox. The first quarter moon is about 90° behind the sun on the ecliptic, where "behind" is said in reference to the daily rotation of the celestial sphere. The sun tends to fall back relative to the fixed stars, so the moon is near the summer solstice.

At sunset in spring, the sun is on the horizon in the west, and the moon, 90° behind, is crossing the meridian while at the summer solstice. It is therefore high in the sky.

We can reason similarly about the fall. In the fall, the sun is near the autumnal equinox, and a first quarter moon, behind the sun on the ecliptic, is near the winter solstice. □

We can imitate the reasoning just given for other phases and seasons. For example, in summer, a full moon is near the winter solstice and thus is low in the sky. You will practice reasoning like this with some exercises. The moon is called "third quarter" a week after it is full.

27.2 *Parallax and Observation*

Where is the moon in the sky? On the one hand this is easy to answer; we look up and see it. In another way, though, this is difficult to answer. The difficulty arises from something called parallax.

Evidence for lunar parallax comes from solar eclipses. On very rare occasions, the moon passes directly between the earth and the sun. When this happens, the sun is partially or even entirely obscured. This is called a solar eclipse.

When a solar eclipse happens, it is visible to relatively few people. An eclipse that is total at one location (the sun completely obscured) will be a partial eclipse at most locations a couple hundred miles away. What does it mean if, at the same time, some people see the moon totally blocking the sun while others do not? It means that the apparent location of the moon in the sky depends not only on time but also on the observer's location. Different observers see the moon in different places—some see it at exactly the same location as the sun, while other do not. The word "parallax" is used to name this effect.

You can experience parallax by looking at something far away while holding one finger in front of your face. Keep looking at the far away object, and close one eye. Then open that eye while closing the other eye, all while continuing to look far away. Your finger seems to change position relative to the far away object. Is your finger to the left of the object, or to the right? It might depend on which eye you use. This is parallax.

The fact that the moon's location in the sky depends on the observer's location on the earth means that we need to take special care when describing the motion of the moon. If we are not careful, we might accidentally involve effects that are due not to the moon itself but instead to the location of the observation.

27.3 Months

One term for measuring time that we have not yet defined is the month. There are of course the various months of the calendar, from 28 to 31 days long. We are interested in months defined using the motion of the moon.

It is straightforward to observe the passage of the moon through its various phases. This yields one definition of month.

Definition 126. *A synodic month is the time from one full moon to the next.*

The synodic month depends on the moon and the sun. Instead of considering the moon relative to the sun, we can also consider it relative to the stars.

This definition is slightly loose. The moon's latitude varies over time, as we will later see, so that it also moves in the direction perpendicular to the ecliptic. A more formal definition would say that the moon returns to the same "right ascension," a term we will not otherwise use.

Definition 127. *A sidereal month is the time it takes for the moon to return to a given place in the celestial sphere.*

Over the course of four weeks, the sun falls back relative to the celestial sphere by almost 28°. This means that the moon needs to travel farther (it is falling back relative to the sun) to arrive at a new synodic month. Thus, the synodic month is longer than the sidereal month.

The synodic and sidereal months are fairly easy to observe, at least approximately. Over a long period of time it is possible to count how many occur in a given number of days. In this way we obtain what are called the mean synodic month and the mean sidereal month. These are averages over long periods of time. The mean synodic month is roughly 29.53 days. The mean sidereal month is roughly 27.32 days.

We need to distinguish between real and mean synodic months, just as we distinguished real and mean days.

There are two more notions of "month" that are worth describing. One of them is in a later section. The other, which we will treat now, involves ideas similar to those used in discussing the equation of time.

It turns out that the moon, which falls back relative to the sun from day to day, does so at a varying rate. Sometimes the moon moves further in the sky along the ecliptic, and sometimes it moves less. The average daily motion is roughly 13°, but the motion varies from day to day. It is not always the same.

It is a remarkable fact that ancient astronomers in Babylon were able to observe the slight variations in the moon's daily motion and to discover that within these variations there is a consistent pattern which repeats about once a month. We are not able to explore the details now, but will simply accept the result.

Definition 128. *The anomalistic month is the amount of time it takes for the moon to pass through its regular variation in speed.*

The anomalistic month is about 27.55 days long. We saw earlier that an epicyclic model for the sun's motion let us account for its variable speed. In our coming epicyclic model for the moon, the anomalistic month will be the period at which the epicycle rotates.

27.4 Epicyclic Model

We will now establish an epicyclic model for the moon. Recall that when we established such a model for the sun, we used accurate knowledge of the lengths of seasons, i.e., the time between solstices and equinoxes. We had such knowledge because people can observe those times fairly accurately. If you had enough time, say a few years, you could start to make those sorts of measurements as well.

With the moon, it is more difficult to make the analogous measurements. The reason is the parallax mentioned earlier. The moon's apparent location in the sky depends on where the observer is on earth.

It is a marvelous fact that we can still track down accurate positions of the moon by elementary means. The way that we do this is through lunar eclipses. During a lunar eclipse, the earth stands between the sun and the moon, so that the moon which would otherwise be full instead is darkened by the shadow of the earth. This means that the moon is exactly opposite the sun, by 180° along the ecliptic. We have a good model for the motion of the sun, so our knowledge of the sun's position, together with the fact that the angle is determined exactly during an eclipse, means that we determine the moon's position as well.

27.4.1 Ancient Lunar Eclipse Data

Three lunar eclipses were seen in close succession—within a couple of years of each other—around 720 BC. These lunar eclipses allow us to nail down fairly accurately the position of the moon at each time. Using this information, we can make a sophisticated geometrical argument and determine the size of the lunar epicycle relative to the average distance from the earth to the moon.

Here are the three lunar eclipses.

Lunar eclipses are seen at night, so we record here the day preceding and the day following the night on which the eclipse was seen.

Lunar Eclipse Dates
March 19–20, 720 BC
March 8–9, 719 BC
September 1–2, 719 BC

The ancient records contain even more information. They include the precise times of the eclipse, the locations where the eclipses were seen, and the extent to which the moon was eclipsed. We can ignore these, since we are making a fairly coarse approximation.

27.4.2 *Lunar Configurations*

To compute the moon's location, we need to know how it moves over time. Here are approximate values of the mean sidereal and anomalistic months.

Sidereal Month	27.32 days
Anomalistic Month	27.55 days

The sidereal month is the amount of time for the moon to return to a fixed star. The anomalistic month is the time for the moon to complete the cycle of its various daily speeds. These times are not the same. This means that the rate at which the epicycle goes about the earth (the sidereal month) differs from the rate at which the epicycle itself rotates (the anomalistic month). The latter is slower. This is depicted in Figure 27.1, in a somewhat exaggerated manner. When the moon is about to return to the same point in the celestial sphere, its anomalous motion is slightly behind (i.e., not yet complete).

Note that the epicycle rotates clockwise while the larger circle rotates counterclockwise.

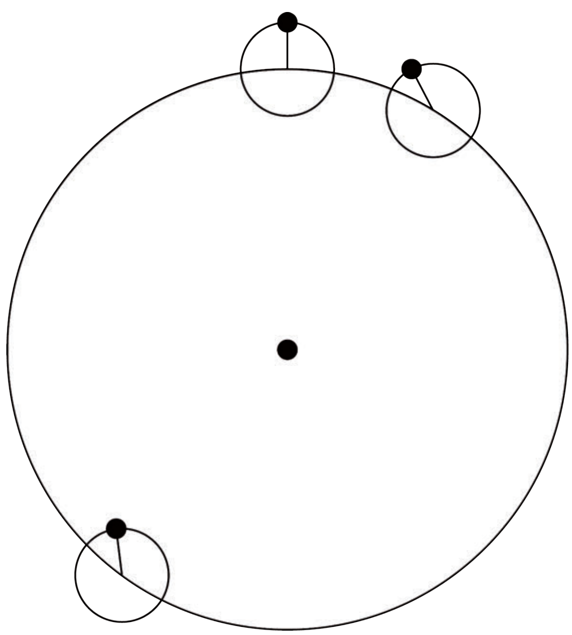


Figure 27.1: Epicycle rate of rotation

The precise times of the eclipses let us know the intervals of time between them, and our knowledge of the sun's motion lets us find how much the sun moves in that span of time. Some details of how this is done are given in an exercise. For now, we will accept the following results of Ptolemy about the sun's motion.

The time from the first to the second eclipse was 354 days and

2.5 hours, in which time the sun moved 349.25°. The time from the second to the third eclipse was 176 days and 20.5 hours, in which time the sun moved 169.5°.

We know that the moon moved by those same amounts in those periods of time, since eclipses occurred, which means that the moon and sun were directly opposite each other, separated by the earth. The epicyclic model divides the moon's motion into two parts. One part is the motion of the whole epicycle around the earth. The other part is the rotation of the epicycle. The first of those parts corresponds to the average sidereal month. The second corresponds to the average anomalistic month (since the rotation of the epicycle provides for the variability in the moon's speed). We will find the size of the epicycle by examining the extent to which the moon's real motion differed from its mean motion (i.e., the motion of the center of the epicycle) in the given time spans.

The moon travels 13.18° per day in the celestial sphere, on average. From this we find that its average motion in 354 days and 2.5 hours is 347.04° more than a whole number of complete cycles (i.e., we ignore complete revolutions of 360°). On the other hand, from the motion of the sun between the first two eclipses, we know that the moon's real motion was 349.25° in that time. That means that the epicycle contributes $349.25 - 347.04$, which is 2.21°, beyond the average motion of the moon between the first and second eclipse.

This value for the moon's mean daily motion is found by dividing 360° by 27.32, the mean sideral month, which is the time taken to travel a total of 360°.

In 176 days and 20.5 hours the average motion of the moon is 170.88°. Thus, the epicycle contributes $170.88 - 169.5$, which is 1.38°, between the second and third eclipses. In this case the epicycle is contributing in the other direction, opposite to the epicycle's motion about the earth. In other words, the moon's observed motion over that period of time is less than the average motion.

The anomalistic month of 27.55 days means that the epicycle rotates at roughly 13.07° per day. That means that the rotation of the epicycle is 308.09° between the first and second eclipse (ignoring the complete revolutions) and that the rotation of the epicycle between the second and third eclipses is roughly 151.43°.

You can review these computations in an exercise.

Figure 27.2 depicts the earth, the moon, and the moon's epicycle at the times of the three eclipses. One thing has been determined roughly. This one thing is the initial angle (at the first eclipse) between the earth, the moon, and the moon's epicycle. Once that is sketched, the configurations at the other eclipses can also be sketched. The reason is that we know the sidereal and anomalistic months and so we know all the subsequent rotations.

In Figure 27.2 the earth is at the center of the circle on which the epicycle rotates. The observer stands on earth, looking outward towards the celestial sphere. By considering the three eclipses in this

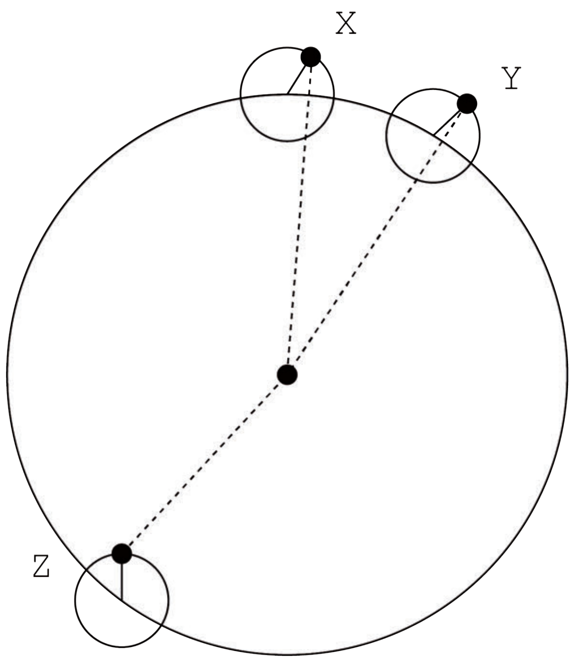


Figure 27.2: Three eclipses

way we can unite the three epicycles into a single diagram, Figure 27.3. The points X, Y, and Z are the locations on the epicycle of the first, second, and third eclipses, respectively. The point E denotes the earth. The diagram is not to scale, and instead exaggerates the relative size of the epicycle (in comparison with the distance to the earth) for ease of use. The points can be interpreted using our earlier computations. We know that the moon moved 2.21° farther than its average motion between the first and second eclipses. Thus, the angle XEY is 2.21°. Similarly, we know that the moon moved 1.38° less than its average motion between the second and third eclipses. Thus, the angle YEZ is 1.38°.

27.4.3 A Euclidean Interlude

Books III and IV of Euclid's *Elements* treat circles and figures drawn in and around them. Our brief treatment of geometry had to overlook these significant topics. To continue reasoning now, though, we must take as given the following result.

Proposition 129 (III.20). *Given an arc on a circle, the angle formed at the center of the circle is twice the angle formed at any point on the circumference.*

This is illustrated in Figure 27.4. The angle ACB is twice the angle ADB. An exercise guides you through the proof.

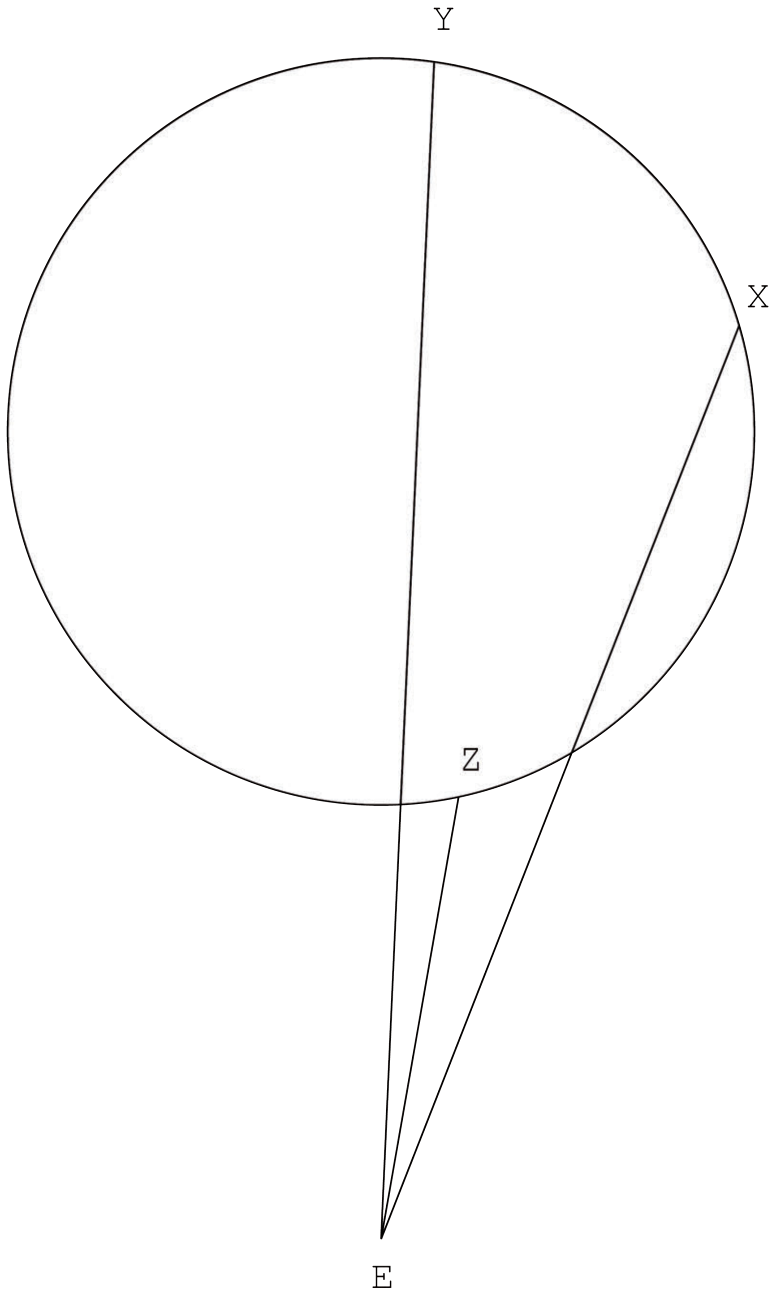


Figure 27.3: Three eclipses, positions on epicycle

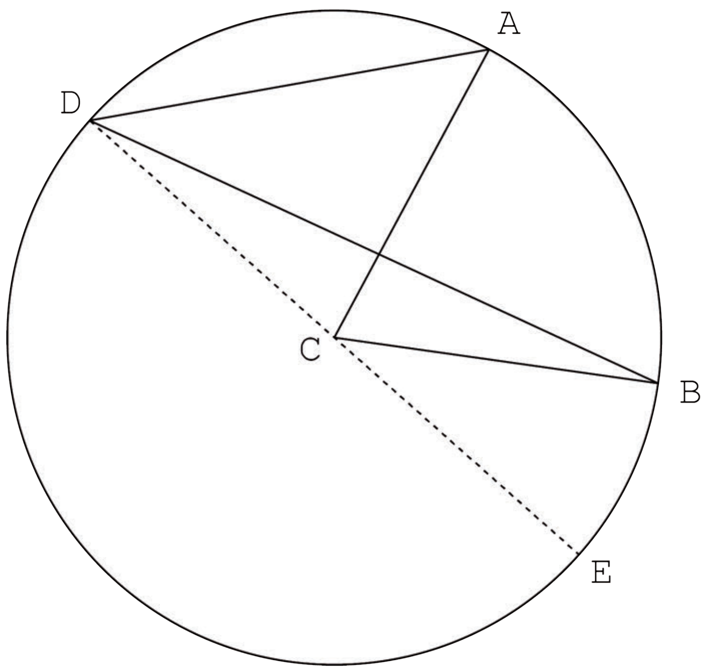


Figure 27.4: Proposition III.20

27.4.4 Computation of Epicycle Radius

Our goal now is to determine the radius of the epicycle. More precisely, we wish to determine the ratio between the radius of the lunar epicycle and the distance from the center of the epicycle to the earth. The strategy we will use is this: we will compute the length of the segment XZ (a chord of the epicycle) in two different ways. One way will involve the ratio between XZ and the distance from the earth to the epicycle. The other way involves the ratio of XZ to the radius of the epicycle. These two ratios will, together, give us the ratio we seek.

In pursuit of our goal, we will rely heavily on our knowledge of chords and the Pythagorean theorem. These two will allow us to grasp many lengths. The argument is not excessively technical, but it can be easy to lose sight of the goal in the details. The skeleton of the proof is given here through a series of lemmas. The proofs of these lemmas are left as guided exercises. The segments are labeled in Figure 27.5. The points N, P, and Q are added so that the angles at N, P, and Q (e.g., angle ENM) are right.

This figure is also not to scale, depicting a large epicycle.

Lemma 130. *Segment MN is approximately* 4.6 *units in length when ME is 120 units in length.*

Lemma 131. *Segment MX is approximately* 11.4 *units in length when ME is 120 units in length.*

Lemma 132. *Segment MP is approximately* 2.9 *units in length when ME is 120 units in length.*

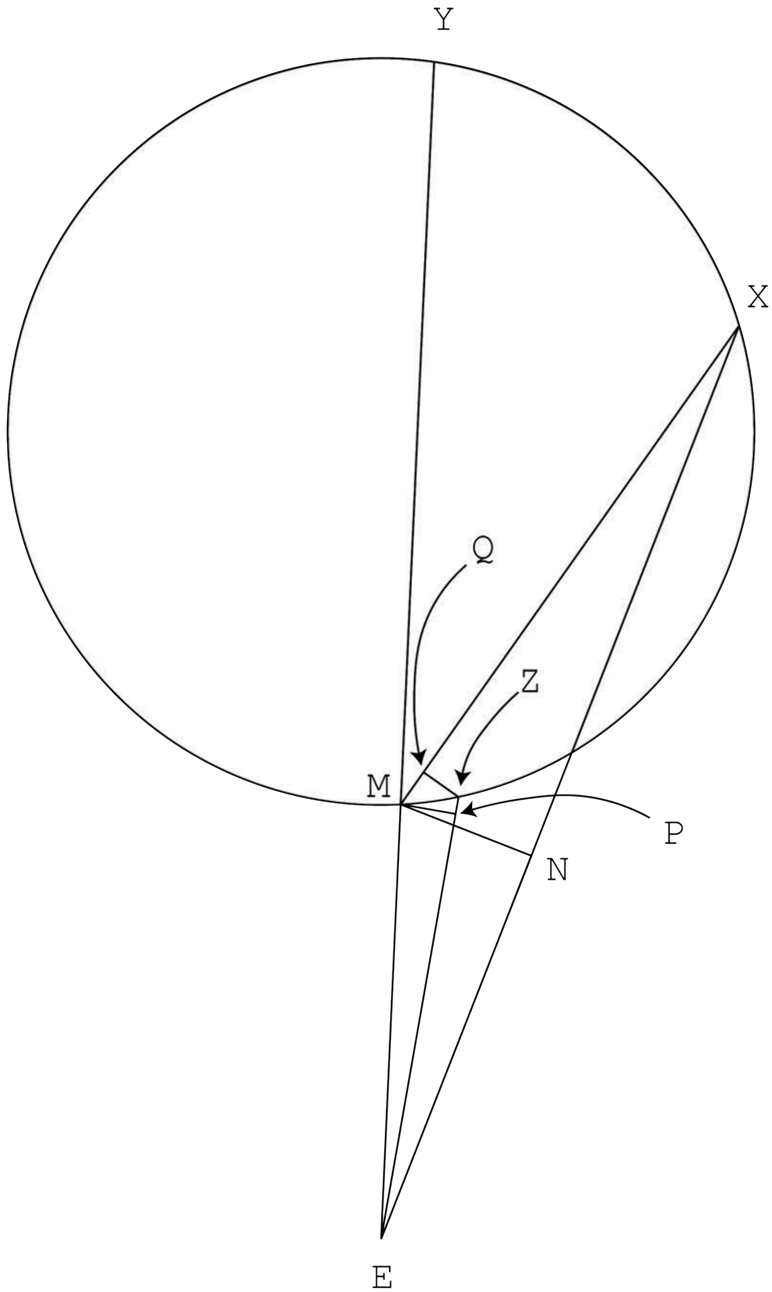


Figure 27.5: Lunar epicycle labeled for computation

Lemma 133. *Segment MZ is approximately* 3 *units in length when ME is 120 units in length.*

Lemma 134. *Segments MQ and QZ are each approximately* 2 *units in length when ME is 120 units in length.*

You can make this lemma more precise. You will also see, later, why such precision is unimportant.

Lemma 135. *Segment XZ is approximately* 9.8 *units in length when ME is 120 units in length.*

The final lemma of that series of lemmas tells us the length of XZ in units involving ME, a segment outside the epicycle. We can also find the length of XZ in units that involve the epicycle, and these two expressions will, together, allow us to infer the relative sizes of the epicycle and the circle about the earth on which it travels.

Proposition 136. *The segment from the earth to the center of the epicycle is roughly* 20 *times the radius of the epicycle.*

Remember that this is a statement about a model.

Proof. By Lemma 135, XZ is roughly 9.8 when ME is 120 units. We can, on the other hand, consider XZ as a chord of the epicycle and use our knowledge of the chord of the corresponding angle. Since the arc YZ on the epicycle is about $151.43°$, and the arc YX is about $51.91°$, we find that the arc XZ is about $99.52°$. Using the chord table, Figure 23.11, this means that the chord XZ is about 91.6 in units in which the epicycle's diameter is 120 units.

We have $99°$ in our chord table. The chord of this slightly larger angle can be estimated from the chord of $99°$. The method is explained in the exercises. For now, it is enough to say that the chord of $99.52°$ is only slightly larger than the chord of $99°$.

The quotient $91.6/9.8$, formed from the measures of XZ according to the two systems of units, is the rescaling factor by which we need to rescale ME to pass from units in which ME is 120 to units in which the epicycle diameter is 120. We compute $120 \times 91.6/9.8$, and conclude that ME is roughly 1120 when units are chosen so that the epicycle has a diameter of 120.

We will now make a useful approximation. The chord YM is not a diameter of the epicycle. As a result, we cannot call EM the distance from the point E to the epicycle, strictly speaking, since there is a slightly shorter segment (one that intersects the epicycle a bit to the left of M in the diagram). Nonetheless, since the arc YM is very large (more than $150°$), we will not err greatly by acting as if YM were a diameter.

The chord of any angle between $150°$ and $180°$ is close to 120.

Supposing YM to be roughly a diameter, we find that the distance from earth to the center of the epicycle is roughly 1180 units (the lemgth of ME plus the radius of the epicycle) when the epicycle radius is 60 units. The numbers 1180 and 60 have a ratio that is approximately $20 : 1$. □

27.5 Anomalous Latitude

The sun and moon both travel along the ecliptic. Why don't we have eclipses every month, then?

It is true that the moon travels along the ecliptic, but this statement can be made more precise. In fact the moon travels along a path that is slightly inclined to the ecliptic. The angle between this circle and the ecliptic is about 5°.

This explains why solar and lunar eclipses are relatively infrequent. The moon tends to be away from the ecliptic by a few degrees. It crosses the ecliptic twice per sidereal month. An eclipse only occurs if the moon crosses the ecliptic at a new or a full moon.

Unlike the ecliptic, which we take to be fixed against the background of the stars, the moon's orbit moves noticeably within the celestial sphere over relatively short periods of time. The two places that the moon's path crosses the ecliptic are called *nodes*. These nodes let us define one more kind of month.

Definition 137. *A draconic month is the time it takes for the moon to return to a node.*

The average draconic month is about 27.21 days. Note that this is shorter than a sidereal month. In consequence, the nodes move westward over time. In other words, the moon passes through the ecliptic at a point slightly ahead (with respect to the rotation of the celestial sphere) of the last crossing it made.

The average draconic month can be obtained by simple, nontechnical method. By observing the moon for a long period of time and noting its change in position in comparison to the ecliptic, we can simply count the passage of days and draconic months and then, after sufficient time, compute an average.

It is possible to incorporate the moon's changing latitude into our epicyclic model. We will not do this, but if you study Ptolemy's *Almagest* in the future, you can learn how it is done.

The word "latitude" is used here in a celestial rather than terrestrial sense. In this setting, we use it to refer to the distance between the ecliptic and an object like the moon.

27.6 Eclipses

Many years ago, people discovered that solar and lunar eclipses occur in certain patterns. We are now in a position to understand one of these patterns.

An eclipse happens when the earth, moon, and sun are all in a single line. This is a delicate configuration, one that is disrupted by a slight change in position. Once such an alignment happens, we can use our knowledge of the various kinds of months to find out possible times for a similar alignment.

Here is an example of this kind of reasoning. Suppose that a solar eclipse happens. This necessarily takes place at a new moon. A solar eclipse will not take place at the next new moon. Why not? The synodic month is about 29.5 days. The draconic month, on the other hand, is about 27.2 days. This means that 27 days after a solar eclipse, the moon is crossing the ecliptic (where it could align with the sun) as a waning crescent, two full days before the new moon. By the time the new moon occurs, two days later, the moon's path has deviated from the ecliptic, and the special alignment that yields an eclipse does not occur.

The moon is said to be "waning" when it is becoming less illuminated, from the perspective of the earth. This is the portion of the month after the full moon and before the new moon.

We have defined four different kinds of months—the synodic, the sidereal, the anomalistic, and the draconic. Let us ignore the sidereal for a moment and consider only the other three. It turns out that each of those months occurs an (essentially) exact number of times in 6585.333 days. We define this period with a special name that possibly originates in a Babylonian word.

Definition 138. *A saros is a period of 6585.333 days.*

Why is it important to consider the saros? The synodic, anomalistic, and draconic months all occur a whole number of times in this period. Thus, after this period passes, the moon is at the same phase as it was at the beginning (it is at a new moon again if it started at new, or at a full moon again if it started at full). In addition, after this period, the moon is at the same point in its speed variation (i.e., in comparison to the mean moon, the center of the epicycle) as it was in the beginning. Finally, the moon is similarly situated with respect to the ecliptic (e.g. at a node, if it was at a node at the beginning).

If an eclipse occurs at a specific time, then there is likely to be an eclipse one saros later. The configuration of the sun, moon, and earth will be almost exactly the same after one saros.

Why did we ignore the sidereal month? We are only interested in the positions of the sun, moon, and earth relative to one another. Since there is a whole number of synodic months in a saros, we also know the position of the sun relative to the moon. To determine the absolute position (i.e., relative to the celestial sphere) involves looking at sidereal months, but this is not relevant if we only want to know whether an eclipse occurs. Sidereal months would tell us where (in the sky) the eclipse occurs.

27.7 Parallax and Computation

At the beginning of this chapter, we saw that lunar parallax can be observed. People who are separated by a fairly small distance have different experiences of a solar eclipse, and thus different judgments about the apparent position of the moon in the celestial sphere.

We can now use geometrical reasoning to infer, using parallax, the distance to the moon. Here is an overview of the method. Using a model of lunar motion like the one we have developed, we can state the true location of the moon with respect to the celestial sphere, i.e., the moon's location if it were to be seen from the center of the earth. We compare this true location with the moon's location as observed from a point on the surface of the earth. The difference between these two, along with our knowledge of the earth's radius and the terrestrial location of the observer, allow us to estimate the distance to the moon.

Ptolemy reports the following. He observed the moon pass the meridian at about 50.9° below the zenith (i.e., slightly less than 40° above the horizon). At this time the true location of the moon, with respect to the earth's center, was about 50° below the zenith. This situation is depicted in Figure 27.6.

These numbers are not exactly what he reported. We have made some small, convenient modifications.

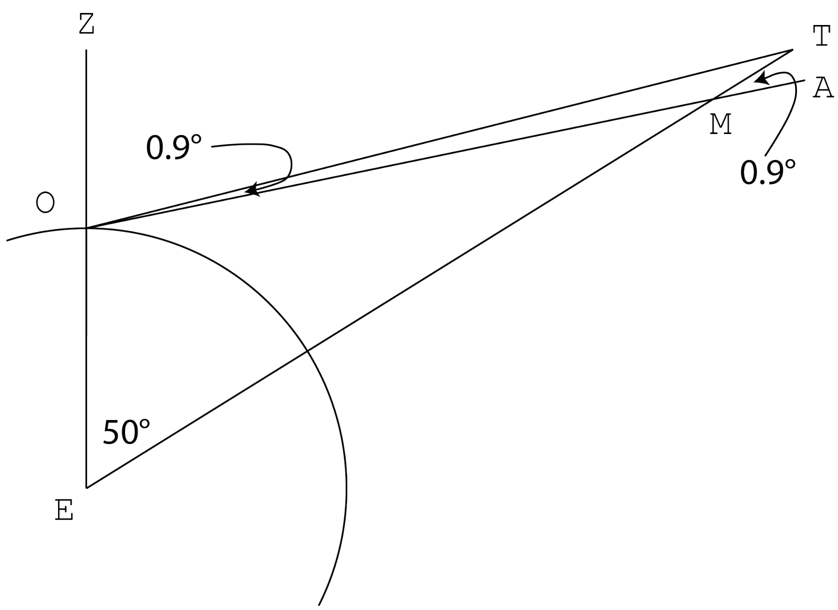


Figure 27.6: Lunar parallax

The point O is the observer, the point E is the center of the earth, the point M is the moon, the point T is the true position of the moon on the celestial sphere (i.e., as seen from the center of the earth) and the point A is the apparent position of the moon on the celestial sphere (i.e., as seen by the observer, Ptolemy). The point Z is the zenith.

According to Ptolemy's lunar model, the angle ZET is $50°$. According to Ptolemy's reported observation, the angle ZOA is $50.9°$.

We want to know angle TOA. We can find it by treating ET and

A model for a computation like this needs to be more complicated than ours. It must take into account the moon's varying latitude, and we did not do that.

OT as if they are parallel. Why? The reason is that the segments ET, OT, and OA are huge in comparison with OM and EM. The celestial sphere has a size so vast that we cannot really compare it to the earth, or to the moon's orbit. As a result, the lines OT and ET, which intersect at T, can be treated as if they were parallel, since near the earth and the moon (far from the celestial sphere) they will appear so. Since they are treated as parallel, the angle ZOT can be treated as if it were $50°$, yielding the conclusion that angle TOA is $0.9°$. Considering alternate interior angles and vertical angles, we find that TMA and OME are also $0.9°$.

Consider two pairs of segments on a piece of paper. One pair of segments is genuinely parallel. The other pair, if you extend them, intersects 100 miles away. Can you tell the difference by looking at them?

From point O produce a line perpendicular to EM, intersecting EM at R. This is depicted in Figure 27.7. The chord of $100°$ is about 92, and the segment OR is 46, half this, when OE (the radius of the earth) is taken to be 120 units. On the other hand, the segment OR is half the chord of $1.8°$ when OM is taken to have 120 units. The chord of $1.8°$ is about 1.9, so that OR is 0.95 when OM is 120 units. The quotient $46/0.95$, about 48, gives the relation in size between the radius of the earth and segment OM (which can be treated as effectively the same as EM).

We have seen this strategy before, where we find the chord of twice the angle in a right angle. See Proposition 114, and in particular Figure 25.5.

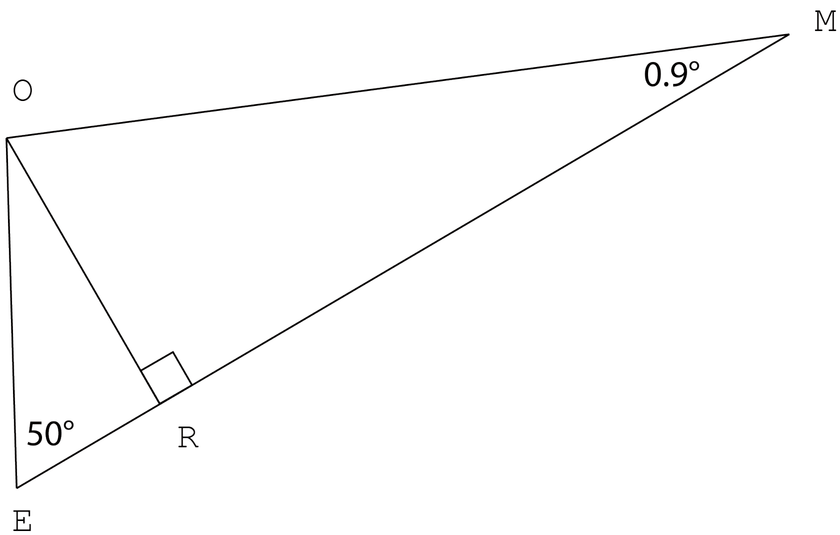


Figure 27.7: Parallax computation

We find that the distance to the moon is something like 50 times the earth's radius, which we know to be about 4000 miles, so we conclude that the distance to the moon is about 200,000 miles. This is a pretty good estimate. The average distance to the moon is about 240,000 miles, as measured by contemporary methods.

Ancient astronomers were in fact able to give much more accurate values than the one we found, values very close to the number accepted today. This is a complicated story. You might want to explore it in the future.

27.8 Exercises

Exercise 1

State for each month of the year whether a full moon in that month at midnight is high in the sky, at a moderate elevation (near the celestial equator), or low in the sky. Explain your reasoning in at least three cases.

Exercise 2

Discuss the appearance of a waxing gibbous moon (between first quarter and full) in October at your latitude. More specifically, state where it rises and where it crosses the meridian.

The moon is said to be "waxing" when it is becoming increasingly illuminated. This is the period of the month after the new moon and before the full moon.

Exercise 3

Discuss the appearance of a third quarter moon in January at your latitude. More specifically, state where it rises and where it crosses the meridian.

Exercise 4

Find an approximate date (being within a month is fine) when each of the following moons rises well to the north of due east.

1. First quarter moon
2. Full moon
3. Third quarter moon

(Hint: where is the sun if the sun rises far to the north of east? Which situations place the moon in that place on the ecliptic?)

Exercise 5

Explain why lunar eclipses do not happen each month.

Exercise 6

Sketch the arrangement of the sun, moon, and earth during a lunar eclipse and during a solar eclipse. Clearly label the diagram for each eclipse type.

Exercise 7

Suppose that the mean sidereal month were 2/3 the anomalistic month. Depict the location of the moon and its epicycle after one

This exercise exaggerates the difference between the sidereal and anomalistic months, but their qualitative relation is the same as that which we observe in reality (the sidereal is shorter than the anomalistic).

(mean) sidereal month. Depict the location of the moon after two sidereal months. Depict the moon after one anomalistic month.

Exercise 8

Prove Proposition 129 (III.20) by reasoning in the following way, with reference to Figure 27.4.

1. Triangle DCB is isosceles (two of its sides are radii) and so two angles in it are the same.

2. Triangle ACD is isosceles (two of its sides are radii) and so two angles in it are the same.

3. Angle ACE is exterior to triangle ACD. Show that it is twice the size of angle ACE.

4. Angle BCE is exterior to triangle DCB. Show that it is twice the size of angle BDC.

5. Conclude that angle ACB is twice the size of angle ADB by considering a difference of angles.

Exercise 9

Here is an explanation of numbers that arose when setting up the lunar epicyclic model. Recall that the anomalistic month is approximately 27.55 days.

1. Divide 360 by 27.55 to obtain the average anomalistic rotation per day. (Round to the hundredths.)

2. Multiply the number from 1. by 354 days 2.5 hours to obtain the rotation of the epicycle between the first and second eclipses. Note that you need to convert 2.5 hours into days, which you can do by dividing by 24 (the number of hours in each day).

3. Multiply the number from 1. by 176 days 20.5 hours to obtain the rotation of the epicycle between the second and third eclipses. Note that you need to convert 20.5 hours into days, which you can do by dividing by 24 (the number of hours in each day).

4. Consider the result of 2. This is the total number of degrees through which the epicycle rotated between the first and second eclipses. Subtract 360 from this number repeatedly, until you obtain a number less than 360. Check that your result matches the one (308.09°) given in the text.

5. Consider the result of 3. This is the total number of degrees through which the epicycle rotated between the second and third eclipses. Subtract 360 from this number repeatedly, until you obtain a number less than 360. Check that your result matches the one (151.43°) given in the text.

Exercise 10

Describe the places on earth at which the moon can sometimes be seen directly overhead. At which of these locations can the sun be seen directly overhead at some point? Explain the difference.

Keep in mind that the moon stays close to the ecliptic, but deviates from it somewhat.

Exercise 11

Make a sketch of the moon's path (relative to the celestial sphere) from one month to the next. Show how the path crosses the ecliptic. Estimate the size of the arc between the two crossings, using the difference between the sidereal month and the draconic month.

Exercise 12

Here are a series of dates (in a non-leap year) on which full moons were observed: January 28, February 27, March 28, April 27, May 26, June 24, July 24. Use these dates to estimate the mean synodic month. You need to count how many total days elapse (remember that calendar months have various lengths), and how many synodic months pass.

Exercise 13

Eclipses (lunar or solar) can occur two weeks apart. Explain how this can happen. What sort of eclipse would each one be? Where would the moon be positioned relative to the node in each case? How much has the sun moved in this time?

Exercise 14

How many years is a saros?

Exercise 15

If you divide 6585.333 (days in a saros) by 29.53 (approximate days in a mean synodic month), you get roughly 223. Check that this is so. Then go the other way. Suppose that one saros contains exactly 223 synodic months. Compute (by hand) the quotient 6585.333/223 to get a refined value for the mean synodic month. Go at least as far as the ten thousandths place (four places after the decimal). After doing this, look up the most accurate value you can find for the mean synodic month.

Exercise 16

How many anomalistic months are in a saros? Draconic months?

Exercise 17

Explore the extent to which the lunar distance computation is sensitive to errors of various kinds. Consider small variations in the parallax, the forecast, and the terrestrial latitude. Which contribute most significantly to a change in the conclusion?

Exercise 18

Pick a date around the present, at random. Suppose that a solar eclipse is seen on that day in southern Europe, around 3pm (in solar time local to the location in Europe). Suppose also that there is a solar eclipse one saros later.

1. What is the date of the second eclipse? (This will depend on the date you chose for the first fictional eclipse.)
2. Where might the second eclipse be visible? (Hint: It will not be visible at the same European location. Will it be visible to the east, or to the west, of the original region of visibility? To answer this, you need to think about where the sun's position after a fraction of a day passes.)

At noon in California, where does an observer in New York see the sun?

Exercise 19

Prove Lemma 130 in the following manner.

1. Angle MEN is the angular difference, as seen from the earth, between the moon at the first eclipse and the moon at the second eclipse, after accounting for the moon's mean motion. What is this angle? It was found in Section 27.4.2. You simply need to track it down.
2. Consider a circle with diameter ME. Use Proposition 129 (III.20) to show that the length of the segment MN can be found as the chord of twice the angle MEN.
3. You need to find the chord of $4.42°$. The chord of $3°$ is 3.15. Chords for small angles increase at a rate of roughly one unit per degree. Since $4.42°$ is bigger than $3°$ by $1.42°$, we will simply add 1.42 to 3.15, to obtain 4.57 as an approximation.

Refer to the chord table Figure 23.11 on page 218 for this chord value, along with those needed in subsequent exercises.

4. (Optional) Here is a way to improve the approximation slightly. Determine the difference between the chord of $3°$ and the chord of $6°$. Call the difference D. Divide D by 3, and then multiply the result by 1.42. Add this to 3.15, the chord of $3°$. In other words, your estimate is $3.15 + \frac{D}{3} \times 1.42$.

This is called linear interpolation. We find the total amount that the chord changes, and then assume that the rate of change is constant in that range.

Exercise 20

Prove Lemma 131 by following these steps.

1. The angle YMX can be found like this. Using the moon's mean anomalous motion, we found that the epicycle rotation from the first to the second eclipse was 308.09° beyond a whole number of complete revolutions. That means that the long way around from X to Y corresponds to 308.09°. The short way, then, is found by subtracting 308.09° from 360°. The vertex M of the angle YMX is at the circumference of the epicycle, not the center, so we find angle YMX by using Proposition 129 (III.20). The angle at the circumference is half the angle at the center, and the latter is known.

2. The angle YMX is an exterior angle of the triangle MEX, and the angle MEX is known. The angles in a triangle are two right angles, so the angle EXM is known. What is it?

3. Consider a circle with XM as diameter, and find MN as a chord of an angle by using Proposition 129 (III.20) as we did in the previous exercise. When you use the chord table, find the nearest value given, and then work away from that by using the approximation of 1 unit per degree (or linear interpolation), as in the previous exercise. This gives the length of MN in units where MX is 120.

As a reminder, suppose that you want to find the chord of 44.5°. You know that the chord of 45° is 45.92, and the angles differ by half a degree, so you can estimate that the chord of 44.5° is 45.42.

4. You should have found, in the previous step, that MN is 48.31 in units in which MX is 120. We now need to rescale to units where ME is 120. Lemma 130 tells us that MN is 4.6 in those other units. To rescale MX (which is 120 in the first units), multiply 120 by 4.6/48.31. This gives MX in units in which ME is 120.

Can you use statements about ratios to justify this method of changing units? Think about the ratios $MN : MX$ and $MN : ME$. We want to know the ratio $MX : ME$. Use Proposition 49 (V.22). To apply that proposition exactly, you need to convert the numerical ratio 120 : 48.31 to a different, but equivalent, form.

Exercise 21

Prove Lemma 132. Reason as you did with Lemma 130. Here is one way to find the chord of 2.76°. We know the chord of 3°. Subtract 0.24 units to arrive at an estimate of the chord of 2.76°.

Once again we are approximating chords of small angles as increasing at roughly one unit per degree.

Exercise 22

Prove Lemma 133.

1. The angle of arc YZ is known (how much did the epicycle rotate, beyond complete revolutions, in the time between the second and third eclipse?). By Proposition 129 (III.20), the angle YMZ is known. What is it?

2. Angle EMP is also known (note that the angle at P is right) so by subtraction we can find angle PMZ.

Keep in mind that EMY is a straight line, i.e., two right angles, meaning 180°.

3. Compute the size of MP in units in which MZ is 120. Use a circle that has MZ as a diameter, and note that the segment MP is opposite the angle MZP in the right triangle MPZ. (Angle MZP can be found since angle PMZ is known and MPZ is right.)

4. Combine the conclusion of the preceding step with the conclusion of Lemma 132 to convert to units in which ME is 120. (Multiply 2.91 by 120 and then divide by the number from the previous step.)

Exercise 23

Prove Lemma 134. Angles YMX and YMZ are known, so angle QMZ is known. There are a couple ways to proceed.

1. (More approximate way) The angle QMZ is close to 45°. We can treat triangle MQZ as if it were an isosceles right triangle. What do we know about the ratio of the diagonal of a square to one of its sides? That ratio is roughly 7 : 5.

2. (More exact way) Consider a circle with MZ as diameter, and use the chord table as we have done in previous parts to find MQ and QZ.

Remember, you can directly check this approximation of the ratio of the diagonal of a square to its side, without any complicated computation. Just take out your compass and straightedge, make a square, and then copy the segments the requisite number of times.

In either case: having found MQ and QZ relative to MZ, now use Lemma 133 to convert to units in which ME is 120.

Exercise 24

Prove Lemma 135. Consider the right triangle XQZ. Its hypotenuse is the thing sought. The segment MX is known, as is MQ, so QX is known too, by subtraction. Furthermore, QZ is known. Then XZ is found using the Pythagorean theorem.

Now you can confirm that our earlier imprecision regarding the lengths of MQ and QZ does not matter. They are small relative to QX and XZ, and so minor adjustments in their values do not play a large role in the final result.

Exercise 25

The argument in Section 27.7 that angle TOA is 0.9° involves a statement about parallel lines and interior angles. Which specific Euclidean proposition is this? Find that proposition in Part I: Geometry, and complete the argument of Section 27.7. Does that proposition rely on Euclid's Fifth Postulate, or not?

Remember that there are two propositions involving these terms, and one is the converse of the other. Be sure that you can identify which one we are using here.

Exercise 26

We found, in Proposition 136, that the moon's epicycle is about one twentieth the size of the average distance of the moon from the earth. Combine this result with the distance we found in Section 27.7 to give the range of distances (least and greatest) between the moon and the earth.

Exercise 27

Look up the currently accepted values of the average distance between the moon and the earth, as well as the least and greatest distances. Compare them to what you found in the previous exercise. Some people might call our numbers fairly accurate. Others might say our numbers are inaccurate. Discuss.

Exercise 28

A better value of the moon's average sidereal motion is $13.176°$ per day (we used $13.18°$ in our computations). A better value of the moon's average anomalistic motion is $13.065°$ per day (we used $13.07°$ in our computations). Investigate how the ratio of Proposition 136 changes with these slight changes in the average daily motions.

Since the daily speeds are multiplied by large numbers (the many days that pass between the eclipses), the tiny changes in the daily speeds accumulate to result in fairly significant changes.

These more accurate speeds will give you a result similar to the one that Ptolemy calculated.

28
Stars, Fixed and Moving

We can see much more in the sky than the sun and moon. This chapter will briefly explain some of the other things we can see.

28.1 Planets

The stars remain in the same positions within the celestial sphere over long periods of time. We do not see them move. There are some things that we do see move, though, and these are called planets.

The planets, like the sun and moon, move along the ecliptic. Their motions, however, are less uniform. One simple number that we can associate to each planet is its *synodic period*.

Definition 139. *The synodic period of a planet is the amount of time it takes for it to return to the same position relative to the sun.*

The use of "synodic" here is like that in "synodic month" which refers to the moon's position in relation to the sun.

The planets we see are divided into two classes. The distinction arises from the extent to which their motion mimics the sun's.

28.1.1 Inferior Planets

The planets Mercury and Venus are called inferior planets. They are only seen near the sun. Since they always appear close to the sun, we see them near sunrise or just after sunset.

Our definition of "synodic period" is imperfect for inferior planets. Can you see why? Mercury and Venus vary from being to the east of the sun and to the west. They do not go "all the way around" like the moon. We need to be more specific about the phrase "position relative to the sun."

Venus is the brightest object we see in the sky after the sun and the moon. It remains near the sun, never getting more than about 45° away. Sometimes it is ahead of the sun (and so visible before dawn), and sometimes it is behind the sun (and so visible after dusk). The synodic period of Venus is about 584 days.

Mercury also remains close to the sun. It is not as bright as Venus. Mercury does not get farther than about 25° from the sun. It appears both ahead of and behind the sun, relative to the daily revolution of the celestial sphere, and moves more rapidly through these positions than Venus does. The synodic period of Mercury is about 116 days.

28.1.2 Superior Planets

The planets Mars, Jupiter, and Saturn are called "superior" planets. Unlike Venus and Mercury, they are not always seen near the sun. Instead, they appear over time at any position in the ecliptic, relative to the sun. Like the moon and the sun, they fall back relative to the fixed stars. Like the moon, they fall back relative to the sun.

Mars has a synodic period of about 780 days and is red. Jupiter has a synodic period of 399 days. Saturn has a synodic period of 378 days. These planets also exhibit a certain non-uniformity known as retrograde motion. While their overall motion is to move back relative to the fixed stars, they move at times in the other direction, advancing relative to the fixed stars.

Retrograde motion will be mentioned again, in the concluding chapter about physics beyond the quadrivium.

28.2 Precession of the Equinoxes

We have taken the ecliptic to be a fixed circle within the celestial sphere. This was a basic feature of the model that we used to describe the motions of the sun and moon, one that made many other things possible. It turns out that this is only an approximation. The ecliptic is not perfectly fixed within the celestial sphere. Instead, over time, the places at which solstices and equinoxes occur move.

It is easier to explain the motion by changing perspective. Rather than thinking of the celestial sphere as fixed, think of the ecliptic as a fixed circle, and the celestial sphere as something moving. Over time, the celestial sphere rotates along the ecliptic. It does so in a way that is roughly the opposite of its daily east-to-west motion. This means that a star that is slightly ahead (relative to rising and setting times) of a point on the ecliptic will, after some time, rise simultaneously with that point, and after even more time it will have fallen behind.

An astronomer named Hipparchus discovered this motion, and estimated that it occurred at a rate of about 1° per century. This is reasonably close to the currently accepted value of 1° in about 72 years.

Hipparchus developed much of the spherical trigonometry that we have used in our study of astronomy.

We distinguished in Section 26.1 between solar and sidereal days. We see now that there are multiple reasonable senses of "year." One, called the "tropical year," involves the sun's return to a solstice or equinox. Another, called the "sidereal year," involves the sun's return to a specific star. Due to precession, these times are not quite the same.

28.3 *Exercises*

Exercise 1

Suppose that Mercury is visible in the morning in early February. What other times that year would you expect to see Mercury in the morning?

This means that Mercury is to the west of the sun, along the ecliptic, so that Mercury has risen before the sun.

Exercise 2

Suppose that Venus is prominent in the evening sky on the summer solstice. When do you next expect to see it prominent in the evening sky?

Exercise 3

When two planets appear close together in the sky, it is called a *conjunction*. Suppose that Jupiter and Saturn appear close together. How long will you wait for the next conjunction?

1. In 399 days, the sun travels roughly 34° beyond a full circle. This is Jupiter's synodic period, meaning that after that amount of time, Jupiter and the sun are the same angle apart from each other. Thus, Jupiter has traveled about 34° along the ecliptic. How much does Jupiter move per day, on average?

For simplicity, we suppose 360° in 365 days, and then one degree per day for the remaining 34 days.

2. Reason as in the previous item, but for Saturn rather than Jupiter. How much does Saturn move per day, on average?
3. Find the difference of the average planetary motions. This is the amount that the distance between the planets increases, on average, per day.
4. How many days does it take for the planets to be 360° apart? (Divide 360 by the difference of the daily motions.)
5. Convert the days from the previous item to years. This gives (roughly) the amount of time between two conjunctions of Jupiter and Saturn.

There are a couple of reasons that this number is rough. One is that our stated synodic periods are not exact values. They can be refined. Another is that retrograde motion, mentioned above, means that there are some complexities in the motions of Jupiter and Saturn. The synodic periods alone do not reflect these motions, which affect the specific times of conjunctions.

Exercise 4

In this exercise, you consider conjunctions between Jupiter and Saturn first qualitatively, and then quantitatively.

1. In the previous exercise, you used a slight overestimate of Jupiter's synodic period, and a slight underestimate of Saturn's. Knowing only that the value (399) for Jupiter was an overestimate and the value for Saturn (378) was an underestimate, is the conjunction period you that found an overestimate, or an underestimate? Explain.

2. More accurate values for the synodic periods are 398.88 (Jupiter) and 378.09 (Saturn). Use these values to compute the time between conjunctions of these two planets. How does it compare to the original value you found? How much do they differ?

Exercise 5

In this exercise, we will determine the time between conjunctions of Mars with the other superior planets. To do so, we must think carefully about the daily rate of motion of Mars through the sky. Recall that the synodic period of Mars is 780 days.

1. Suppose that the sun and Mars are at the same point on the ecliptic. For concreteness, let us say that this occurs on June 21, the summer solstice that year. Draw a sketch of this, viewing the ecliptic from above, with the earth at the center.

The specific date is simply to make subsequent descriptions easier.

2. After one year, the sun has returned to the summer solstice. Mars has fallen back relative to the stars. Show that Mars must have fallen back on the ecliptic by more than 180° by reasoning like this. If it had fallen back by less than 180°, then Mars would travel less than 360° in two years, which means that the Sun would overtake Mars in the second year, so that the synodic period would be less than 730 days.

730 is 2×365.

3. We conclude from the previous part that Mars must have traveled a bit more than 180° in the first year. In other words, Mars is closer to the spring equinox than the autumnal equinox. Sketch this.

4. Let D denote the number of degrees that Mars travels in one year. Conclude, from the synodic period of Mars and the reasoning of the preceding parts, that $2D$ is roughly 410. Find D, the degrees traveled by Mars per year.

410 is 360 (the number of degrees in a complete revolution) plus 50 (the additional distance traveled by the sun in the fifty days by which 780 days exceeds two years).

5. Use the number in the preceding part to find the daily angular motion of Mars.

Now reason as in the previous exercise, using the daily motion of Mars along with the values found earlier for Jupiter and Saturn. Consider the difference in the daily angular motions, and determine how long it takes for these differences to add up to 360°.

Exercise 6

Discuss the periods of conjunctions between superior and inferior planets. Keep in mind that inferior planets are always fairly close to the sun. You do not need to be numerically precise.

Exercise 7

How long does it take for the celestial sphere and the solstices and equinoxes to line up again in the same way? This means, how much time does it take for precession to go through 360°?

Exercise 8

The north celestial pole is marked fairly accurately by a star called Polaris. Discuss where Polaris will appear after a few thousand years. Consider the precession of the equinoxes. If you would like, do the following.

1. Lie on your back, with your head pointing due south and your feet to the north. Where would you see Polaris? Point at it.

We are presuming the exercise is done in the northern hemisphere.

2. Take a plate or some other flat disc in your hands. While still lying on your back, hold the disc so that it marks out the celestial equator. Tip it the right amount, based on your terrestrial latitude. Turn the disc to mimic the daily motion of the celestial sphere. (This means, turn it counterclockwise.)

3. Suppose, for simplicity, that you are completing the exercise on the summer solstice (even if you are not). Picture the ecliptic. How does it relate to the celestial equator? Tilt the disc so that now it marks out the ecliptic.

4. Turn the (ecliptic-marking) disc clockwise. That mimics the precession of the equinoxes. How does Polaris move, relative to the solstice, if the celestial sphere turns like that? To determine how much to turn the disc, use your computation of how long it takes for the precession of the equinoxes to go through 360°.

Exercise 9

The Tropic of Cancer is the northernmost latitude at which the sun appears directly overhead. This corresponds to the obliquity of the ecliptic, 23.5°. The word "Cancer" is the name of a constellation, a collection of stars. The ecliptic passes through the constellation Cancer. At the time the Tropic of Cancer was named, about 2000 years ago, the sun was at the summer solstice while in the constellation Cancer. Due to the precession of the equinoxes, the solstice moves over time to different places in the sky. When will the term "Tropic of Cancer" have two essentially opposite meanings? Use the value 1° per 72 years for the precession of the equinoxes.

In other words, when will the winter solstice occur when the sun is in the constellation Cancer?

Exercise 10

Determine whether the tropical year or the sidereal year is longer. By how much? Use the given value of precession, and the approximation that the sun travels 1° per day, to approximate the time difference.

After one tropical year, the sun returns to the exact same place on the ecliptic. What has happened to the background stars?

Part V

Beyond the Quadrivium

Multiplicati sunt super numerum.

29
Physics

The music and astronomy we have studied are a small, but important, part of what we know about the world. Here are some further steps you might take as you continue to learn. In each case it is possible, and important, to see how your new knowledge relates to what you learned previously.

29.1 New Models in Astronomy

Our astronomical models were based on circles. It is possible to understand circles by elementary means. The chord computations at the heart of our method relied on the Pythagorean theorem, properties of regular polygons, and a few other things that can be grasped without too much difficulty.

Our astronomical models were also based on observations from the earth. We spoke of the appearance of things as we stand on the ground. We can readily pass between mathematical features of our model and the things that we see when looking up at the sky.

Models of astronomical motion can be made much more sophisticated than ours. Here are three steps of that process. The three steps are associated with the significant scientists Copernicus, Kepler, and Newton.

Copernicus proposed that the earth and the planets should be modeled as moving about the sun. This was an idea that some Greek thinkers had also considered many years before, but which had not gained wide acceptance. We mentioned retrograde motion in passing, when discussing the superior planets. One appealing feature of the Copernican model is that it simplifies the account of retrograde motion. Copernicus used circles and epicycles, like Ptolemy did, but their center was to be the sun rather than the earth.

Kepler, working about 50 years after Copernicus, proposed to explain planetary motion about the sun using three laws. Two of these laws are important modifications of simple features of our models. One is that ellipses, rather than circles, are used for the paths of the planets. Another is that the rate at which the planets move along their paths is taken to be non-uniform.

Recall that the epicyclic and eccentric models are equivalent. The eccentric model involves a uniform speed (about a point other than the earth) and so the epicyclic model involves a uniform speed as well.

Newton followed by roughly another half-century. He was an outstanding scientist and mathematician who made many significant contributions. One of his insights was that the three laws of Kepler could arise from a single principle with wide explanatory power. This general principle is Newton's law of universal gravitation. It accounts for Kepler's laws and unites our experience of weight on earth with the motions of the sun, moon, and planets in the sky.

29.2 *Polyphony and Temperament*

Our study of music was simple. It led to the Pythagorean tuning system, pleasantly intelligible and uniform. This system of tuning is reasonable for music consisting of successive single pitches. Such music can be called *monophonic*, meaning that it consists of only one sound at a given time. In contrast to monophony, music can also by *polyphonic*, involving multiple simultaneous sounds. Polyphony has arisen in various places and times. Gregorian chant, the monophonic music which had been the dominant ecclesiastical music of the West, was supplemented with polyphonic forms in the 12th and 13th centuries. Development has continued since then, leading to the great variety of church and secular music we hear today.

In studying the Pythagorean system, we found that two whole tones yield a somewhat awkward ratio. The composition of two ratios $9 : 8$ results in the ratio $81 : 64$, which exceeds the simpler, superparticular ratio 5 : 4 in a way that is discernible to the untrained ear. The latter ratio, 5 : 4, is better than 81 : 64 for many musical purposes, especially when notes sound simultaneously. This means that people seek new tuning systems, especially for polyphonic music.

One method, called "well temperament," is mentioned in the title of Johann Sebastian Bach's *The Well-Tempered Clavier*. Pianos today are tuned in what is called "equal temperament."

It turns out that no perfect solution exists. Each tuning system, like the Pythagorean one, makes some compromises. Some intervals are cleaner and purer, others are more muddied. There is not room here to discuss this at length. The essential idea, though, is the one that we saw in Proposition 98, that no number of fifths is equal to an exact number of octaves. That proposition has its roots in the prime numbers themselves.

You might also have completed the exercise to prove Proposition 100, which generalizes the result about fifths to arbitrary superparticular ratios.

We feel vibration in our throats when we speak, and we see it in the string that we pluck. This movement is wave-like. As a result, the study of music can be understood as part of a more general study of wave motion. Waves are pervasive, and they occur in more places than the sea. In the 19th century, scientists came to think of light as a kind of wave.

29.3 *Experiment and Induction*

We hear sounds and see the sun, moon, and stars without effort. The scientific investigations of music and astronomy in this book were limited to these simple, natural experiences. It is possible to go beyond such experiences, though, and such efforts have directed scientific research for many years. It is often necessary to do experiments before reasoning in the ways that we have here. Experiments give us new experiences of order in the world, order which can be subject to careful investigation with mathematical language.

One thing that we hope to glean from experiments is new principles. We might also hope to clarify and refine existing principles. This is somewhat different from the way that we reasoned here about music and the heavens; our principles were quite natural, just as our geometrical postulates were. They did not need a great deal of justification. Scientists who study things that are unfamiliar and distant from our ordinary experience must find ways to become acquainted with the basic phenomena in order to find suitable principles.

The term "deduction" applies to the sort of reasoning that goes into a mathematical proof. We have relied on deductive reasoning throughout our study of the quadrivium. A different name, "induction," is given to the act by which we proceed from experience (including experiments) to basic principles. A greater emphasis on experiment means a greater emphasis on induction.

29.4 *Immediacy and Dependence*

The musical and astronomical phenomena of this book are largely part of your immediate experience. Nothing stood between you and the thing being discussed. You heard the musical sounds yourself. You saw the sun rising and setting.

Astronomy presents some significant exceptions to the general immediacy of the quadrivium. We know about eclipses through reports given by other people, rather than through our own direct experience. We also rely on other people when we determine that the sun's motion throughout the sky each year is uniform over all the earth, varying from about 23.5° north of the celestial equator to 23.5° south. We could not observe this in all locations on earth, even if we had many years to try.

As scientific study considers more complex phenomena, it loses the immediacy of elementary music and astronomy. There are two related kinds of mediation. One kind of mediation is the sort that we see in the eclipse reports. When we read about an eclipse, we learn about something that we understand and can imagine, even if we do not encounter it directly. We can choose to trust that the report is accurate and then proceed in our reasoning, acting as if the encounter had been our own. A second kind of mediation involves the use of instruments. You were able to build a monochord, or at least you understand the way the thing is built. More complicated study requires us to use instruments that other people build. These instruments then give us information about something that we encounter directly. A simple example of this is a thermometer. You can touch something, like water, and determine whether it is fairly warm or cool. You do not assign a number to this. If you buy a thermometer, you can get a number. While this number deals with something that you experience directly (the warmth or coolness of the water) the way that the number arises is hidden from you. You do not know how the thermometer itself works. It was built by someone else, whom you must trust.

In both cases—remote phenomena and instrumental mediation—we see our dependence on other people. Our scientific knowledge becomes inextricably bound up in many kinds of trust: trust that the reports other people give are true, trust that the instruments other people build are reliable. Our scientific knowledge rises from a foundation with two parts: the experience and principles that we grasp clearly on our own, and our faith in a common activity.

30 Mathematics

Mathematics, both in the ancient world and today, includes much more than our geometry and arithmetic. Euclid's *Elements* alone is already a substantial work; you have seen only a small portion of its contents. Return to Euclid to learn more ancient mathematics. Here are some modern developments.

30.1 Real Numbers

We use numbers in a variety of ways. We count individual items. We also measure lengths, weights, and volumes. In the first case the use of number is clear; it involves a natural unit. When we wish to know how many rocks are in a pile, we have a clear idea of what is meant by "rock" and we proceed by simply counting them. For extended objects, the use of number is not so immediate. Instead, we make a choice of a unit. This could be something like inches or meters, or we could also make a comparison to some other object, such as when we say that one plot of land is twice as large as another.

One significant choice of unit that we made during our study was when we developed the chord table. We chose units in which the circle had a diameter of 120 units. This was a free choice, but it was also convenient, since it means that the chord and the corresponding angle (in degrees) are roughly the same, at least for small angles.

Try to remember: for which angle is the chord and the angle measure (in degrees) exactly the same? Think about hexagons.

Some ratios of segments are not realized as ratios of natural numbers. This leads to what we now call "real numbers." The basic idea of the real numbers is that there are segments of "any possible size." We can imagine a segment being just a bit longer, or shorter, and given two segments of different sizes we can imagine one whose length is between the other two. We can fix a specific segment as a unit. Then, given any other segment, we can consider the ratio of that segment to the unit segment. Such a ratio is a real number.

Find a proposition in Part III, on Music, where we proved this.

This is a sketch and not a proper definition. The definition of real number involves ratios of natural numbers and not of segments.

Calculus is a significant area of mathematics that treats curves and models of motion. Calculus is rooted in the real numbers, in the notion that we can pass freely between numbered and extended things by way of ratio. If you have the opportunity to study calculus, pay close attention to statements about continuity, or ideas of filling gaps. In these statements, you are encountering the real numbers in the foundations of calculus. Be sure to ask your teacher questions to see why the real numbers matter in what you are doing.

One important instance in which real numbers matter is called the *Intermediate Value Theorem*.

30.2 Functions

The chord table shows a relationship between two different kinds of things, chords and angles. We used geometrical reasoning to find some specific values, and continuing with such reasoning would allow us to go even further to produce more entries in the table. It is possible to think not only of individual entries, but also to consider the relationship as a whole. This whole relationship is called a "function."

Functions pervade contemporary mathematics. They do a number of things. One is to reflect ideas about causality. Another is to make geometrical things more numerical and subject to computation.

Functions are a good way to capture notions of cause and sequential behavior. In the chord table, we can think of the angle as an "input" and the chord of the arc as the "output." The choice of an angle, an input, yields a definite other thing, the chord, thought of as an output. In many circumstances we have an idea of the ways that one thing depends on other things, and when this is so we can use functions profitably for modeling.

Functions are often used in physical reasoning, with time as the input. We can count seconds, hours, minutes, days, or years, and in this way associate numbers to periods of time. When something changes over time, such as the height of a falling rock above the earth or the temperature of a cup of hot water sitting on a counter, we can produce a model in which time is the input to a function. The output is distance (for the rock) or temperature (for the cup of water). If we can produce a model for the behavior over time with a single function, that single function captures something uniform and intelligible in the behavior, just as the notion of arc and chord in a circle unifies the collection of all the entries in the chord table.

Functions are also useful as a way to make geometrical things computable. This happens by relating functions to geometry through the notion of a graph. A graph is a curve that arises by considering the inputs and outputs of a function. Collecting all these pairs and displaying them in a plane, we obtain the function's graph. We can also go the other way, starting with something geometrical and considering the function whose graph is that geometrical curve.

The basic questions of calculus involve functions and their graphs. One question is about how a graph (which might be curving) can be approximated by a line. This leads to what is called the derivative. Another question is about the size of the region enclosed by a graph, and leads to what is called the definite integral.

Here is a warning about functions. When you work with functions in your ongoing mathematical education, you might be tempted to speak of them as being in some way incomplete (and your book or your teacher might do so too). This often happens when we talk about "evaluating" functions, i.e., determining which output corresponds to a given input. Keep the example of the chord table in mind. The relation between arc and chord is clearly independent of our thinking. The functional relation, the chord table, is entirely present at the outset. Our "discovery" of certain values brings about a change only in our own knowledge, not in the mathematical thing. If you can keep in mind that functions are stable, complete, unchanging, you will avoid some errors that give students difficulty with functions.

30.3 *Trigonometry and Fourier Series*

The word "trigonometry" is built on terms for triangles ("trigon") and measurement ("metry"). The chord table is a kind of trigonometry: consider the isosceles triangle formed by a chord of a circle along with the two radii terminating in the chord's endpoints. The chord table tells us the length of the side that is not a radius.

Mathematical developments after Ptolemy have emphasized slightly different relations for measurement in triangles. These are the 'sine' and 'cosine' functions. The sine and cosine functions are closely related to the chord function. They involve a different standard choice of unit, one in which the radius of the circle is the unit.

Let α be an angle. See if you can describe the relationship between the chord of α and the sine of an angle related to α. You will need to keep in mind our choice of unit for the chord table, and you should also make use of Proposition 129 (III.20). What can you say about the sine of half of α?

If you study trigonometry, you will learn various trigonometric identities. These identities are relationships like the ones we used to find chords from other chords (such as finding the chord of the sum of two angles). They are proved using the Pythagorean theorem and other fundamental geometrical properties you now understand.

Recall Ptolemy's theorem (Theorem 108) and the discussion that followed.

The trigonometric functions enable us to model periodic behavior. This is useful in music, astronomy, and many other areas. Recall that *timbre* refers to the character of a sound. It varies among voices and instruments even when the pitch being produced is the same. One way to think about timbre is as different waves coming together to form one single, complex motion. Combining trigonometric functions, which are like models of waves, gives something called a Fourier series. We can also think of the epicycles in Ptolemy's astronomy as a way of combining simple periodic motions to create a single complex motion.

30.4 Non-Euclidean Geometry

Euclidean geometry is founded upon five postulates. The last of the postulates, which involves two lines intersected by a transverse line, is the most complex. People wondered whether this final postulate could be proved from the other postulates. If this were the case, it could be omitted from the postulates and instead included as a proposition.

After many years of study, mathematicians were able to show that the fifth postulate cannot be proved from the other four. It is independent of them. This justifies Euclid's foundation of geometry. There is nothing redundant in it.

Mathematicians became interested in studying what could be proved in a logical system in which Euclid's fifth postulate does not hold. This study is called non-Euclidean geometry. It seems strange, at first, but becomes more tangible by thinking about the points and lines as lying on curving surfaces, like spheres and saddles.

30.5 Algebra

In arithmetic it was useful to give names to natural numbers. We began statements with phrases like "let m be a natural number." In such a statement the thing that is named by m is fairly definite. We can imagine two copies of m objects, and we denote that as $2m$. We can imagine the product of m with itself, and this is written as m^2. As mathematics has developed, people have come to work more and more with such symbolic expressions, in a way that often obscures the thing designated by the symbol. This kind of mathematics is called algebra.

We stated and proved a theorem about the existence of solutions of linear Diophantine equations. The standard formulation of this theorem is more algebraic. It asserts that something involving certain symbols is equal to some other thing. When you see such an equality being asserted, you should take the time to think carefully about what is being named by each thing in the symbolic expression.

Algebra is related to geometry as well. Symbols like x and y are used to refer to certain segments in a plane. Complicated equations involving x and y determine curves in the plane (some of these curves are graphs of functions, mentioned earlier). The study of how equations involving symbolic expressions are related to geometric shapes is called "analytic geometry."

31 Philosophy

The word *quadrivium* was created by Boethius, a Roman author who sought to preserve and hand on Greek learning. Boethius's studies of arithmetic and music guided our exploration of those disciplines. At the beginning of his work on arithmetic, Boethius says that this four-part study, the quadrivium, prepares us to seek the highest, unchanging things.

31.1 Foundations of Mathematics

To make definitions at the beginning of our study of geometry, we needed to use words that already had meaning. We could not begin from nothing. Instead, we were confident that certain simple words have meanings that are shared by all people, and we relied on that universal intelligibility to build up the rest of our study. Similarly, in arithmetic, we did not make an entire set of axioms explicit. Instead, we proceeded directly, trusting that all people know what natural numbers are.

Many people have thought about and discussed the building blocks of mathematics and how we come to know them. One basic question is whether people create mathematics or discover it. It is best to avoid answering this question too quickly. Instead, learn a good deal of mathematics first. Your study of the quadrivium is a good start.

One significant recent development in the foundations of mathematics is the use of abstract collections, called "sets." By making sets the foundation for mathematics, it is possible to unify geometry and arithmetic to a great extent. There is a cost, however, which is that a proper study of sets is rather complicated. It is not sufficient to accept a simple notion of sets in ordinary language without substantial qualification.

To learn why, find out about something called "Russell's paradox."

31.2 Logic and Mathematics

When we talk about mathematical proofs, are we doing mathematics? This is a challenging question, and one that points to a broader one. To what extent is logic distinct from mathematics?

It is clear that geometry and arithmetic are not, themselves, the whole of logical thinking. The power of mathematical abstraction is very broad, and so mathematics includes all kinds of order (physical or otherwise) that we discern in the world. Much logic can therefore be included within mathematics. To determine the boundaries of the two is a philosophical question.

When we set out postulates, as in geometry, and work within them, we are doing mathematics. It is also possible to make the postulates themselves the object of a kind of mathematical study. This is the sort of thing that is done when people show that Euclid's fifth postulate is independent of the others. Such investigation can be called "metamathematical."

31.3 Models in Physics

Ptolemy showed that the eccentric and epicyclic models for the sun lead to the same account of its motion. We also examined that equivalence. This comparison shows that the specific mathematical formulation is subordinate to the phenomenon itself. There is no reason, in what we see physically, to prefer one model to the other, and thus we should be slow to imagine that the elements of the model, such as the epicycle, themselves have physical meaning.

See Proposition 115.

Physics abounds in mathematical models, and it is often difficult to distinguish between fundamental, sensible things being modeled and mathematical artifacts of the model. One example of this, which plays a key role in most introductions to calculus, is instantaneous velocity. Time and distance are clearly discernible by us, in a direct fashion. By choosing units we can make them numerical, and their quotient gives us the notion of "average speed." By introducing a functional model of the motion, it is possible to arrive at what is called the "instantaneous velocity." This is an answer to the question "How fast is it going right now?" It is important to realize that the instantaneous velocity is not directly accessible to our senses and is not measurable with an instrument. It is instead something that arises from a model of motion, from something mathematical added on to the tangible, physical thing.

The line between the thing being modeled and the model itself becomes especially faint when we explore things that are furthest from our ordinary experience, like atoms and their constituents. We then deal almost exclusively with mathematical entities. Interpreting what these mean physically is a significant philosophical problem.

31.4 Analogy

Mathematics and physics are a small part of the world that we describe and investigate through speech. Many who study the quadrivium will not go on to professional studies of scientific disciplines. For such people, though, the effort is not wasted. Mathematics is a powerful, enduring source of analogies.

An analogy is a comparison that explains or clarifies one of the items compared. Things, or relationships between things, can be said to be alike. To do this is to draw an analogy. Explanation by way of analogy does not mean that the things being compared are exactly the same. Instead, it means that they are similar in an important manner, one worth highlighting.

Mathematical reasoning is especially clear, compelling, and universal. It is a good source for analogies. In this way, mathematical study bears fruit in other kinds of knowing. Here are two examples.

31.4.1 Justice and Means

Virtue is something between excess and defect. It is intermediate between these two. Consider the specific virtue of justice. There are a number of situations in which we can strive to be just. One is when rectifying a wrong that one person has done to another. Another is when sharing an investment or business venture with partners. A third is in buying and selling. There is a single thing, justice, that should be present in each case.

To understand more of what justice is, it is possible to draw an analogy to the three means: the arithmetic, the geometric, and the harmonic. Each is a number intermediate between other numbers. One makes differences the same, another involves ratio, and the third is mixture of the other two. This is something like justice in retribution, investment, and exchange. Retributive justice makes equal, in some way, the difference that arose from the injustice. Just sharing of investment profits means returning to the various parties in a way proportional to their contribution. Justice in exchange is both adversarial and collaborative. We wish for a good price, but will only arrive at one through cooperative behavior that keeps our counterpart involved.

The philosopher Aristotle talks about means in this way in his work called the *Nicomachean Ethics*.

When we say that there is an analogy between the kinds of justice and the three means, we do not say that they are exactly the same. We cannot hope for perfect justice by quantification and accounting.

The purpose of making the analogy is to shed light on the thing that we grasp obscurely and to provoke new questions through the comparison.

31.4.2 Ratio and Analogy

The theory of ratio developed by Eudoxus and transmitted in Book V of Euclid's *Elements* is a profound contribution to human thought. It is also something that ties together the various strands of the quadrivium. Through ratio we understand similar figures in geometry, multiplicative relations in arithmetic, consonant pitches in music, and arcs on the celestial sphere.

These many appearances convince us that ratio is pervasive and powerful. We can put it to one final use, beyond mathematics, to examine analogy itself. What is analogy?

Analogy is like equality of ratio. A single ratio arises as a relation between two things, such as segments. Two things in a different domain, like numbers, can also have a ratio to one another. We can determine whether or not the two ratios are the same. Suppose that they are the same. What have we concluded? There is some common way that the two segments, and the two numbers, are related internally. That equivalence of their internal relations is an external relation, one that connects their limited domains to a larger sphere. In the end, it unites them to the whole.

General Index

Index of Euclidean Propositions