Chapter

9

TRIGONOMETRY

The word "trigonometry" literally means "the measure of triangles". However, from its inception what we today call "trigonometry" has been utilized to deal with a wide variety of problems. From its roots in Greece around 300 B.C. until the 17th century, developments in trigonometry were driven by problems in astronomy involving the locations and relative motions of the sun, moon, planets, and stars, as well as by earth-bound problems of surveying, navigation, map making, and construction. These applications required the measurement of angles, the lengths of line segments and circular arcs, and areas of geometric figures in the plane and on the surface of a sphere. As a result, plane and spherical trigonometry developed simultaneously with many fruitful connections between them, but with spherical trigonometry preeminent.

In the 17th century, mathematicians and scientists began to realize that the algebraic, graphic, and analytic properties of the trigonometric functions could be exploited to solve problems in new areas of pure and applied mathematics. For example, trigonometry played a basic role in the development of coordinate geometry and calculus. These developments were stimulated by such problems as modeling periodic behavior, representing and approximating numbers and functions, analyzing wave phenomena and mechanical vibrations and motion problems, and representing and analyzing curves and surfaces. In this setting, the connections between plane trigonometry and the analysis of functions predominated. Though spherical trigonometry continued to play an important supporting role in certain applications, plane trigonometry's role became larger.

The primary emphasis of trigonometry, as it is currently taught in high school and college, is on plane trigonometry, and that is the focus of our discussion in this chapter. Because trigonometric content is typically presented in several different courses in high school and college, we provide here a more global perspective of the basic principles and concepts of plane trigonometry as well as a sample of the diversity of its applications. We also discuss in more detail the fascinating historical and conceptual evolution of trigonometry. This historical and evolutionary perspective is essential to understanding and teaching trigonometry within the broader context of mathematics and science.

Unit 9.1 Angle Measure and the Trigonometric Ratios

9.1.1 Angle measure and arc length

In this section, we discuss the common units for measuring angles and arcs. You may wish to refer to the discussions of angles and rotations in Section 7.2.2 and of arcs in Section 8.2.4.

Units of angle and arc measure

Degree measure, introduced by the Babylonians between the years 2000 and 1600 B.C., is not only the oldest surviving system for measuring angles, but also the most familiar. A **degree** can be defined as the measure of an angle $\angle ACB$ that corresponds to a rotation of the initial side $\overrightarrow{CA}$ to the terminal side $\overrightarrow{CB}$ of magnitude $\frac{1}{360}$ of a revolution around C. More briefly,

$$1 \text{ degree} = \frac{1}{360} \text{ of a revolution.}$$

The other common measure is the **radian**. This word was introduced around 1870 by Thomas Muir and James Thompson, Sr., to stand for "radial angle". Given an angle $\angle ACB$, and a circle of radius r centered at the vertex C of the angle, **the radian measure of** $\angle$**ACB** is $\frac{s_r}{r}$, where s_r is the length of the arc with central angle $\angle ACB$ on the given circle.

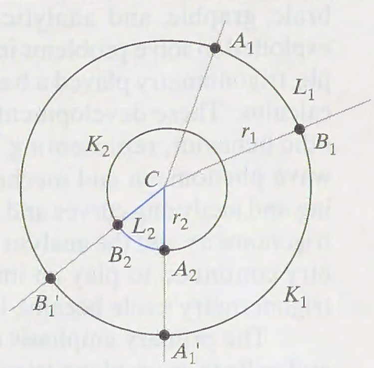
Because the circumference of a circle of radius r is $2\pi r$, this definition implies that

$$2\pi \text{ radians} = 1 \text{ revolution.}$$

This definition raises a mathematical question, "Does the radian measure of an angle depend on the radius r of the given circle in that definition?" The following result shows that it does not.

Theorem 9.1 **(Arc-to-Radius Similarity Principle):** Suppose two arcs have the same measure on circles with radii r_1 and r_2 and a common center C. If the arcs have lengths L_1 and L_2, then $\frac{L_1}{r_1} = \frac{L_2}{r_2}$.

Figure 1



Proof: Let the arcs $\widehat{A_1B_1}$ and $\widehat{A_2B_2}$ have the same measure in the circles K_1 and K_2, both centered at C, as shown in Figure 1. Then a suitable rotation around C maps the central angle $\angle A_1CB_1$ onto the congruent central angle $\angle A_2CB_2$ and the arc $\widehat{A_1B_1}$ onto an arc $\widehat{A_1'B_1'}$ on K_1. The length of $\widehat{A_1'B_1'}$ is equal to L_1, the length of $\widehat{A_1B_1}$ because a rotation is a congruence. Then a size change transformation centered at C of magnitude $\frac{r_2}{r_1}$ maps the circle K_1 onto the circle K_2 and image arc $\widehat{A_1'B_1'}$ on K_1 onto the arc $\widehat{A_2'B_2'}$ with length L_2. This size change is a similarity transformation, so $L_2 = \frac{r_2}{r_1}L_1$. Consequently, $\frac{L_1}{r_1} = \frac{L_2}{r_2}$. ∎

Degree measure is directly related to radian measure by the radian-degree conversion formula

$$\text{measure of angle in radians} = \tfrac{\pi}{180} \cdot (\text{measure of angle in degrees}),$$

because 2π radians $= 360°$.

The formula for the length s of an arc of radius r with a central angle of measure θ in radians is especially simple and elegant. The fraction $\frac{\theta}{2\pi}$ represents how much of the circle's circumference has been traversed. Multiply that by the circumference $2\pi r$ and we arrive at the following theorem.

Theorem 9.2 Let s be the length of an arc of a central angle with measure θ radians in a circle with radius r. Then $s = r\theta$.

If the central angle θ is measured in degrees rather than radians, we can still compute the length s of the arc but the formula is not as elegant. Using the radian-degree conversion formula, we obtain this corollary.

Corollary: Let s be the length of an arc of a central angle with measure θ degrees in a circle of radius r. Then $s = \frac{r\pi\theta}{180}$.

In many applications, either Theorem 9.2 or its corollary can be used, as we choose. However, the simplicity of the formula in radians leads to a corresponding simplicity in analytic formulas for the trigonometric functions. For example, the following familiar formulas from calculus are valid only if θ is measured in radians.

$$\frac{d}{d\theta}(\sin\theta) = \cos\theta \qquad \frac{d}{d\theta}(\cos\theta) = -\sin\theta$$

$$\sin\theta = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \cdots + \frac{(-1)^k\theta^{2k+1}}{(2k+1)!} + \cdots$$

$$\cos\theta = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \cdots + \frac{(-1)^k\theta^{2k}}{(2k)!} + \cdots$$

These formulas are more complicated if θ is measured in degrees (see Problem 8).

Other angle measure systems

Although radian and degree measure are the most commonly used systems for angle measure, other units are useful in certain contexts. The following are illustrations of such systems.

In **grad measure**, which is used in some engineering contexts,

$$1 \text{ grad} = \frac{1}{400} \text{ of a revolution.}$$

Grad measure is better suited to the decimal system than degree measure because a right angle has a measure of 100 grads, a straight angle is 200 grads and a full revolution is 400 grads. The formulas

$$\text{measure of angle in radians} = \frac{\pi}{200} \cdot (\text{measure of angle in grads})$$

$$\text{measure of angle in degrees} = \frac{9}{10} \cdot (\text{measure of angle in grads})$$

relate grad measure to degree and radian measure.

In **mil measure**, which was developed for use in some military contexts,

$$1 \text{ mil} = \frac{1}{6400} \text{ of a revolution.}$$

The apparently odd choice of the fraction $\frac{1}{6400}$ is explained by the calculation

$$\frac{2\pi}{6400} = .00098175 \approx \frac{1}{1000}.$$

Thus, an angle of 1 mil subtends an arc of length very nearly 1 yard at a distance of 1000 yards. This approximation makes it possible to use quick mental arithmetic to calculate the distance to objects whose size is known.

9.1.1 Problems

1. What is the radius of a circle for which the radian measure of an arc equals its length?

2. What is the length of a 212° arc in a circle of radius 14?

3. In a circle of radius 2, suppose a chord $\overline{AB}$ has length 1.

a. What are the degree measures of the major and minor arcs $\widehat{AB}$ in that circle?

b. Generalize part **a**.

4. Unique among the measure systems mentioned in this section, the degree contains subunits of minutes and seconds that are often used in contexts such as latitude and longitude and surveying. Specifically, $1° = 60$ minutes, denoted $60'$, and 1 minute = 60 seconds, denoted $60''$.

a. Convert $37°6'52''$ to a decimal number of degrees.

b. Convert $37°6'52''$ to radians.

c. Find $\sin^{-1} 0.4$ to the nearest second.

5. Consider a watch with an hour hand, a minute hand, and a second hand. In a time interval of x seconds, what is the measure of the arc traversed by each of these hands? Give an answer in degrees and an answer in radians.

6. Find the grad and mil measures of the angles with degree measure 30°, 45°, and 60°.

7. An observer sees a tank in the distance and measures the angle of elevation from ground level to the top of the tank to be 2 mils. If he knows that this type of tank is 9 ft tall, approximately how far away is the tank?

8. The following three calculus formulas require the measure of the angle θ to be in radians.

i. $\frac{d}{d\theta}(\sin\theta) = \cos\theta$

ii. $\lim_{\theta\to 0} \frac{\sin\theta}{\theta} = 1$

iii. $\sin\theta = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \cdots + \frac{(-1)^k\theta^{2k+1}}{(2k+1)!} + \cdots$

a. Find the corresponding formulas for degree measure.

b. Find the corresponding formulas for grad measure.

9. The radian measure of an angle θ is defined in a calculus book[1] as follows: Given an angle $\angle ACB$, and a circle of radius r centered at the vertex C of the angle, the radian measure of $\angle ACB$ is $\frac{2S_r}{r^2}$, where S_r is the area of the sector with central angle $\angle ACB$ on the given circle. Explain why this definition is equivalent to the definition given in this section.

10. Prove the converse to Theorem 9.1.

9.1.2 The trigonometric ratios

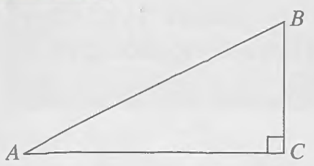
The trigonometric ratios for the sine, cosine, and tangent are so important that they are introduced in textbooks at grade levels from before high school to college. Before the appearance of hand-held calculators, tables of their values from 0° to 90° appeared as appendices in the backs of these textbooks.

Trigonometric ratios of acute angles

Recall that these ratios are defined for acute angles as follows: Let ΔABC be a triangle with right angle at C, as shown in Figure 2.

[1] Apostol, Tom. *Calculus*, Volume 1 (New York, NY: Blaisdell, 1964).

Figure 2



The **sine of the acute angle A**, abbreviated **sin A**, is defined to be

$$\sin A = \frac{BC}{AB} = \frac{\text{length of opposite side}}{\text{length of hypotenuse}}.$$

Similarly, the **cosine of the acute angle A**, abbreviated **cos A**, is defined as

$$\cos A = \frac{AC}{AB} = \frac{\text{length of adjacent side}}{\text{length of hypotenuse}}.$$

The remaining four trigonometric ratios of different sides in a right triangle, the **tangent**, the **cotangent**, the **secant**, and the **cosecant** of angle A are abbreviated and defined for the ΔABC as

$$\tan A = \frac{BC}{AC}, \quad \cot A = \frac{AC}{BC}, \quad \sec A = \frac{AB}{AC}, \quad \csc A = \frac{AB}{BC}.$$

It is straightforward to verify that the value of any one of these trigonometric ratios for a given ΔABC determines the values of the remaining five trigonometric ratios.

Question: If angle A is acute and $\sin A = k$, give the values of the other five trigonometric ratios in terms of k.

From the triangle congruence theorems applied to right triangles, all side lengths and all angle measures for a given right triangle ΔABC are determined if the given information about ΔABC includes either the lengths of two sides (by SAS or SsA, or by the Pythagorean Theorem), or the length of one side and the measure of one of the acute angles (by ASA or AAS). Determining these lengths and angle measures is called **solving the triangle**. The following example is a very simple illustration of this point. More substantial applications of this procedure are discussed in Section 9.1.3.

EXAMPLE 1 Suppose that ΔABC is a right triangle with $\angle C$ as the right angle, and suppose that $AC = 10$ ft and $BC = 13$ ft. Solve the triangle. That is, find the measures of the other sides and angles of the triangle.

Solution By the Pythagorean Theorem,

$$AB = \sqrt{AC^2 + BC^2} = \sqrt{269}\text{ ft} \approx 16.4\text{ ft}.$$

Now any of the trigonometric ratios can be used to determine an angle. We choose to use the tangent.

$$\tan \angle A = \frac{BC}{AC} = 1.3, \quad \text{so} \quad m\angle A \approx \tan^{-1} 1.3 \approx 52.4°.$$

From this, $m\angle B \approx 90° - 52.4° = 37.6°$.

Similarity and the trigonometric ratios

The answers to the question above demonstrate that the value of any one of the six trigonometric ratios determines the values of the other five for a given right triangle. However, to apply these ratios to solve *any* right triangle, we need the following theorem. It lets us use the *measure of the angle* as opposed to the angle itself as the argument for the sine and other trigonometric ratios. Although it is a consequence of the Fundamental Property of Similarity mentioned in Section 8.2.1, it can also be proved without appealing to that result.

Theorem 9.3 **(Right Triangle Similarity Principle):** Let ΔABC be a right triangle with its right angle at C. Then any ratio of two side lengths of ΔABC determines (1) all other ratios of two corresponding side lengths and (2) both of the corresponding acute angles for any triangle $\Delta A'B'C'$ that is similar to ΔABC.

Proof:

(1) We noted above that the value of any one of the six trigonometric ratios for any given triangle determines the values of the other five for that triangle.

(2) Suppose that $\Delta ABC \sim \Delta A'B'C'$. Also suppose that the ratio $\frac{BC}{AC}$ is known. Because these triangles are similar, there is a constant k such that

$$A'C' = k \cdot AC \qquad A'B' = k \cdot AB \qquad B'C' = k \cdot BC.$$

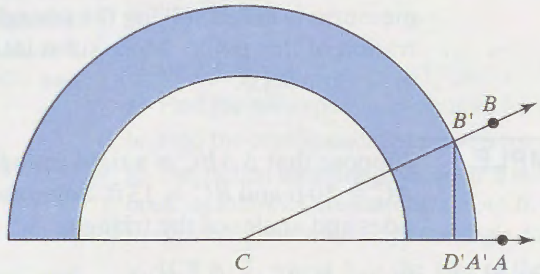
$$\text{Consequently, } \tan A' = \frac{B'C'}{A'C'} = \frac{k \cdot BC}{k \cdot AC} = \frac{BC}{AC} = \tan A.$$

This determines $m\angle A$ and $m\angle A'$, and $\angle B$ and $\angle B'$ are their complements. (If another ratio of sides were given, we would have used a different ratio in place of the tangent ratio.) ┘

Theorem 9.3 enables us to define the sine of the *measure* of the acute angle A to equal the ratio defined above for any angle with that measure.

Now we connect the measures of arcs to the measures of angles. Recall the usual procedure for measuring an angle with a protractor. We place the protractor so that its center point is at the vertex C of the angle $\angle ACB$ and so that the ray $\overrightarrow{CA}$ is along the straightedge of the protractor, as shown in Figure 3.

Figure 3

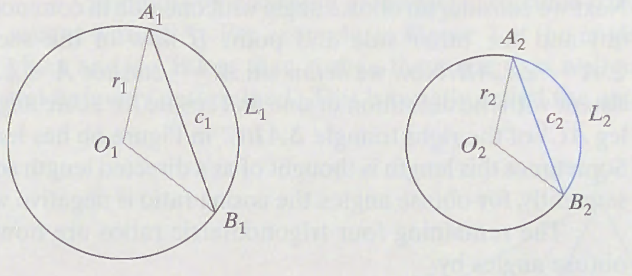


We then read the measure of the angle in degrees from the marks on the arc $A'B'$ on the protractor. Thus, we assign the degree measure to $\angle ACB$ not by a measurement at C, but by a measurement of the length of an arc with center C at some distance from C. Also, note from Figure 3 that the trigonometric ratios for the angle $\angle ACB$ can be computed for the right triangle $\Delta D'CB'$. What we need to show is that the ratio of the length of the leg $\overline{B'D'}$ to the length of the arc $\overset{\frown}{A'B'}$ is independent of the size of our protractor. The following result, which is a consequence of Theorems 8.16 and 8.17, provides one way to obtain the numerical correspondence between acute angles and their measures. It is an analogue to Theorem 9.1.

Theorem 9.4 **(Arc-to-Chord Similarity Principle):** Suppose two arcs have the same measure, and chords with lengths c_1 and c_2. If the arcs have lengths L_1 and L_2, then $\frac{L_1}{c_1} = \frac{L_2}{c_2}$.

Proof: By Theorem 8.16, two arcs have the same measure if and only if they are similar. Suppose arcs $\overset{\frown}{A_1B_1}$ and $\overset{\frown}{A_2B_2}$ are similar in circles of radii r_1 and r_2, respectively, as shown in Figure 4. Then $\frac{r_2}{r_1}$ is the magnitude of the similarity transformation mapping circle O_1 onto circle O_2. So the ratio of the lengths of the chords, $\frac{A_2B_2}{A_1B_1} = \frac{r_2}{r_1}$.

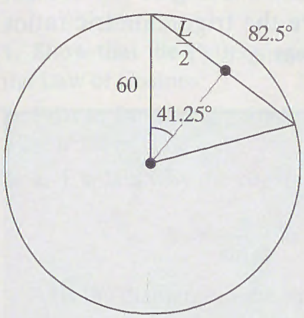
Figure 4



From the definition of arc measure, the central angles O_1 and O_2 of arcs $\overline{A_1B_1}$ and $\overline{A_2B_2}$ have the same measure. Call that measure θ. By Theorem 9.2, the length of $\overline{A_1B_1} = r_1\theta$ and the length of $\overline{A_2B_2} = r_2\theta$. Consequently, $\theta = \frac{\text{length of } \overline{A_1B_1}}{r_1} = \frac{\text{length of } \overline{A_2B_2}}{r_2}$. Now, by SAS Similarity, $\Delta A_1O_1B_1 \sim \Delta A_2O_2B_2$. Thus $\frac{r_1}{r_2} = \frac{c_1}{c_2}$, and the theorem follows. ■

In the historical evolution of trigonometry, in order to obtain measures of angles and arcs, and lengths of sides of figures, Hipparchus and Ptolemy computed tables of the *chord lengths* for a circle of fixed radius for central angles having measures between 0° and 180°. Hipparchus computed these values for multiples of 7.5°. We do not know the radius of the circle he used. By the time of Ptolemy, the Greeks were using a sexagesimal (based on 60) numeration system, so Ptolemy used a circle with radius 60. He was able to calculate these measures for multiples of 0.5°.[2]

Figure 5



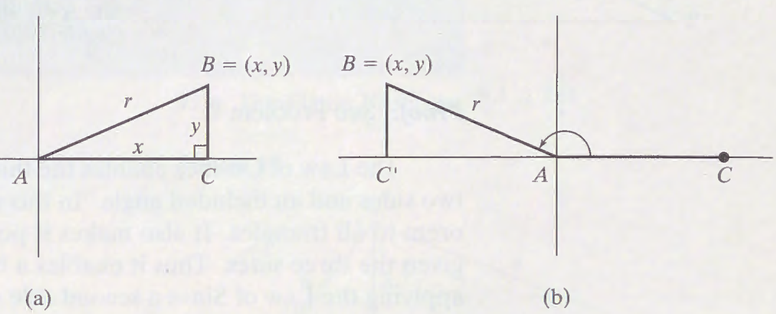
Here is an example of how Ptolemy's table of chords relates to today's values of sines. Suppose an arc has measure 82.5° in a circle of radius 60 and we want to know the length L of its chord. Then the triangle formed by a radius of the circle to an endpoint of the arc and to the midpoint of the chord of that arc (Figure 5) has a central angle with measure 41.25°. So $\frac{\frac{L}{2}}{60} = \sin 41.25°$. From this we see that $L = 120 \sin 41.25°$. In general, the chord lengths in these ancient tables for a given arc measure are 120 times the values of sines of half that arc measure.

Trigonometric ratios of obtuse angles

So far, we have defined the trigonometric ratios for the acute angles of a right triangle. To solve all triangles, trigonometric ratios of obtuse angles are also needed.

The usual method for making the transition for trigonometric ratios from acute to obtuse angles is to begin by placing the acute angle A on rectangular coordinate axes, as shown in Figure 6a. In this position, we see that $\cos A = \frac{x}{r}$ and $\sin A = \frac{y}{r}$, where x, y, and r are the lengths of sides of the right triangle. Also, most significantly, x and y are the coordinates of the point B.

Figure 6



[2] See Asger Aaboe, *Episodes from the Early History of Mathematics*, p. 112ff.

Next we consider an obtuse angle with one side in common with the acute angle (Figure 6b) and the other side and point B now in the second quadrant. Here again $\angle A = \angle CAB$. Now we *define* $\sin A = \frac{y}{r}$ and $\cos A = \frac{x}{r}$. This definition is clearly consistent with the definition of sine and cosine for acute angles. Notice, however, that the leg $\overline{AC'}$ of the right triangle $\Delta ABC'$ in Figure 6b has length $-x$ (since x is negative). Sometimes this length is thought of as a directed length so that it can be negative. Consequently, for obtuse angles, the cosine ratio is negative while the sine ratio is positive.

The remaining four trigonometric ratios are now defined for both acute and obtuse angles by

$$\tan A = \frac{y}{x}, \quad \cot A = \frac{x}{y}, \quad \sec A = \frac{r}{x}, \quad \text{and} \quad \csc A = \frac{r}{y}.$$

Because x is negative while y and r are positive for an obtuse angle A, the ratios $\tan A$, $\cot A$, and $\sec A$ are negative while $\csc A$ is positive.

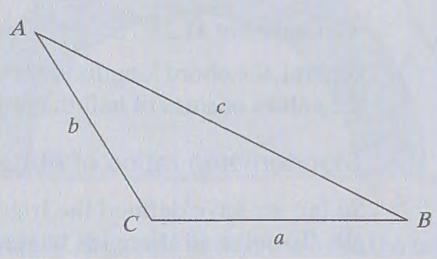
To obtain $\sin A$ and $\cos A$ when A is a right angle, we can adapt either Figure 6a or Figure 6b to the situation where $x = 0$ and $y > 0$. Then B is on the positive ray of the y-axis, and $\angle CAB$ is a right angle. Since $y = r$, $\sin A = 1$, $\cos A = 0$, $\tan A$ is undefined, and so on.

The solution of oblique triangles

In Section 9.1.3, we discuss several examples in which we solve triangles that are not right triangles. The important tools in these solutions are the trigonometric ratios and two theorems, the *Law of Cosines* and the *Law of Sines*.

Theorem 9.5 **(Law of Cosines):** For any ΔABC,

$$c^2 = a^2 + b^2 - 2ab\cos C.$$



Proof: See Problem 6. ∎

Theorem 9.6 **(Law of Sines):** For any ΔABC,

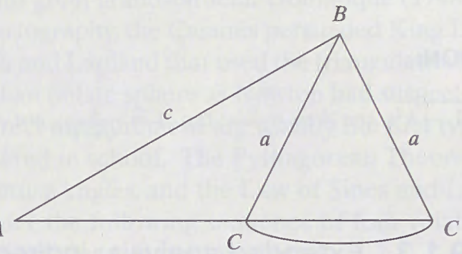
$$\frac{\sin A}{a} = \frac{\sin B}{b} = \frac{\sin C}{c}.$$

Proof: See Problem 7. ∎

The Law of Cosines enables the third side of a triangle to be determined given two sides and an included angle. In this regard, it generalizes the Pythagorean Theorem to all triangles. It also makes it possible to determine any angle of a triangle given the three sides. Thus it enables a triangle to be solved given SAS or SSS. By applying the Law of Sines a second side of a triangle can be found given two angles and a side, and also a second angle can be found given two sides and an angle. So it enables a triangle to be solved given AAS, ASA, or SSA.

Care must be exercised in the last case, since triangles are not necessarily congruent with SSA. For example, in Figure 7, if the angle A and the sides c and a are given and if a is less than c, then the angle C as well as the side length $b = AC$ are not uniquely determined. This is usually called the *ambiguous case for SSA*.

Figure 7



However, if $a \geq c$, then angles C and B are uniquely determined. This is the case of SsA congruence discussed in Section 7.4.1.

9.1.2 Problems

1. Show that the Pythagorean Theorem is a special case of the Law of Cosines.

2. Prove: In ΔABC, angle C is obtuse if and only if $a^2 + b^2 < c^2$.

3. a. Explain why the constant K in the Law of Sines

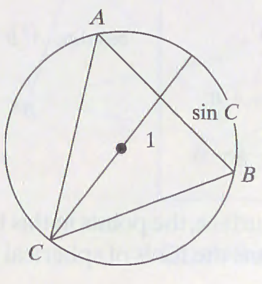
$$K = \frac{a}{\sin A} = \frac{b}{\sin B} = \frac{c}{\sin C}$$

is the diameter of the circumcircle of ΔABC. (*Hint*: If S is the circumscribed circle of the triangle ΔABC, construct a second triangle $\Delta A'B'C'$ in which $A = A'$, $\overline{AB'}$ is a diameter of S, and $\angle B'AC' \cong \angle BAC$.)

b. Use the result of part a to conclude that if ΔABC is inscribed in a circle of diameter 1, then $a = \sin A$, $b = \sin B$, and $c = \sin C$.

4. Explain why the following is an equivalent definition of the sine of an angle: The sine of an angle between 0° and 180° is equal to the length of the chord of the arc subtended by inscribing the angle ($\angle ACB$ in Figure 8) in a circle of diameter 1. (This definition is very closely related to the definition Hipparchus and Ptolemy used to construct the first sine tables.)

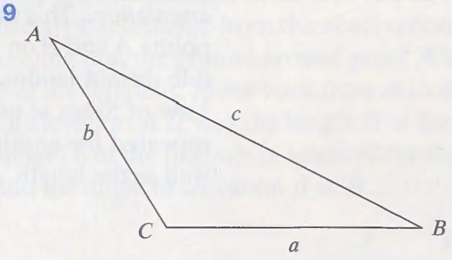
Figure 8



5. Show that Ptolemy's procedure gives tables like sine tables. Let ΔABC be a right triangle inscribed in a circle with diameter $AB = 120$. Prove: If L_{2A} is the length of the chord (from his table of chords) corresponding to the central angle with measure $2A$, then $\sin A = \frac{L_{2A}}{120}$.

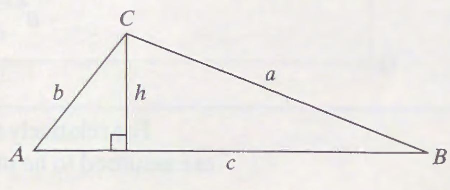
6. Prove the Law of Cosines using the following procedure. Place the triangle ΔABC from Figure 9 so that C is at (0, 0) and side $\overline{CB}$ lies along the positive x-axis. Identify the coordinates of the vertices A and B in terms of the side lengths a, b and the sine and cosine of C. Then apply the distance formula to compute the square c^2 of the length of the side $\overline{AB}$ and simplify the result.

Figure 9



7. a. Use Figure 10. Prove: $\frac{\sin A}{a} = \frac{\sin B}{b}$.

Figure 10



b. Explain how one can go from the result of part **a** to the full statement of the Law of Sines.

8. Suppose sides a and b and $m\angle A$ of ΔABC are given.

a. Prove that there is exactly one triangle satisfying these conditions if $a \geq b$.

b. For what values of $a < b$ is there exactly one triangle satisfying these conditions?

c. When is there no triangle satisfying these conditions?

ANSWER TO QUESTION:

$\csc A = \frac{1}{k}$; $\cos A = \sqrt{1-k^2}$; $\sec A = \frac{1}{\sqrt{1-k^2}}$; $\tan A = \frac{k}{\sqrt{1-k^2}}$; $\cot A = \frac{\sqrt{1-k^2}}{k}$

9.1.3 Extended analysis: indirect measurement problems

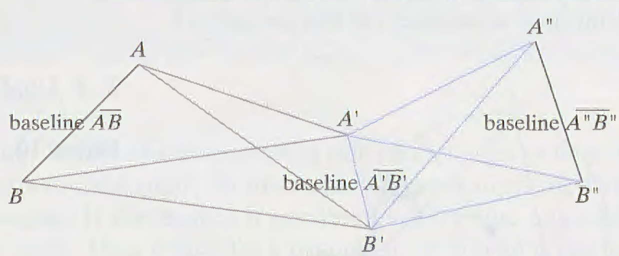
The need to determine distances, angles, and arcs that could not be measured directly led to the study of trigonometry in ancient Egypt and Babylonia. The Rhind Papyrus, a scroll of mathematical problems written in Egypt about four thousand years ago, includes several problems (Problems 56–60) that deal with measurements of angles and distances related to the pyramids. Babylonian astronomers recorded the motions of the planets on the celestial sphere, an imaginary sphere centered at Earth whose surface they believed contained all the stars and planets.

Although these ancient measurements were recorded with meticulous care and detail, they were observational rather than analytical. The development of tables by Hipparchus and Ptolemy and the development of the basic theorems of plane and spherical trigonometry took place between the second century B.C. and the second century A.D. This made it possible to use trigonometry to calculate indirect measurements.

Triangulation

Ferdinand Magellan's circumnavigation of the earth (1519–1522) provided the first direct proof that the earth is roughly spherical. It prompted questions such as whether the earth was actually an oblate sphere flattened at the poles, and stimulated interest in the development of accurate maps of continents and oceans on the earth's surface. As a result, a number of geodetic surveys were conducted based on the *method of triangulation.* This method begins with a baseline segment of known length joining two points A and B on the earth's surface that are a known distance apart. For a given visible distant landmark, A' in Figure 11, $\angle A'AB$ and $\angle A'BA$ are measured, and the Law of Sines is used to compute the distances AA' and BA'. This process can be repeated for another visible landmark B' to compute the distances AB' and BB' as well as the length $A'B'$ of a new baseline $\overline{A'B'}$, and so on.

Figure 11

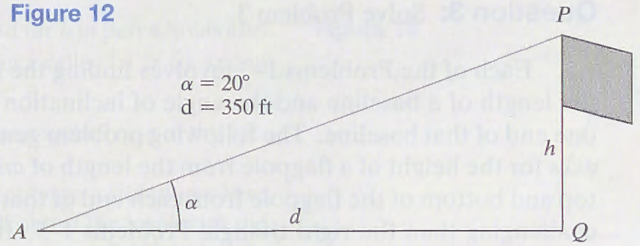


For relatively small regions of the earth's surface, the points in this triangulation grid are assumed to lie in a plane, but for larger regions the tools of spherical trgonometry are needed. Using the triangulation method, Abbé Jean Picard (1620–1682) started with a baseline consisting of a 7-mile stretch of road from Paris to Fontainebleau, and completed a detailed survey of a region extending to a section of the French coastline. This survey was completed for all of France by Giovanni Cassini (1625–1712), an Italian astronomer and cartographer, his son Jacques (1677–1756), his grandson Cesar Francois (1714–1784), and his great grandson Jean Dominique (1748–1848). In the course of these successes in cartography, the Cassinis persuaded King Louis XV to authorize two expeditions to Peru and Lapland that used the triangulation method to determine that the earth is indeed an oblate sphere as Newton had suspected.

Problems in indirect measurement are usually the first types of applications of trigonometry encountered in school. The Pythagorean Theorem, the trigonometric ratios for acute and obtuse angles, and the Law of Sines and Law of Cosines are all that is needed. Consider the following sequence of four related and progressively more complicated indirect measurement problems.

Problem 1: We want to determine the height h of a certain flagpole $\overline{QP}$ (see Figure 12) and we cannot measure its height by climbing it. Instead, we measure $\angle PAQ$, the angle of elevation to the top of the flagpole from a point A on level ground 350 feet from the base of the pole, and find that it is 20°. Find the height h of the flagpole.

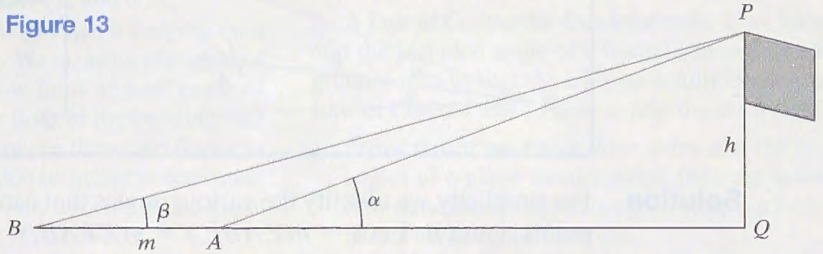
Figure 12



Question 1: Solve Problem 1. Then generalize your solution by solving the problem when the length from the base of the flagpole is d and the angle of elevation is α.

Problem 2: Suppose that the flagpole is located on the other side of a stream so that it is not possible to measure directly the distance from the observation point A to the base Q of the flagpole. Assume that the ground around point A is level and at the same level as the base Q of the flagpole. Move back from A along $\overrightarrow{QA}$ to point B and measure the angle of elevation at B, and the length m of the baseline $\overline{AB}$ (see Figure 13). Find the height h of the flagpole in terms of the distance m, the angle of elevation α at A, and the angle of elevation β at B.

Figure 13

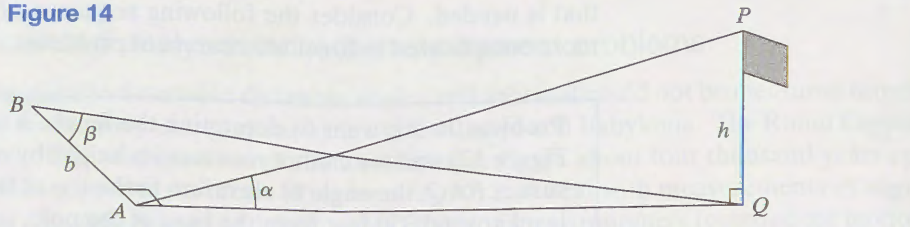


Question 2: Solve Problem 2.

Problem 2 assumes that the measured baseline $\overline{AB}$ is collinear with a line to the base of the flagpole. But in some situations the easiest baseline to measure may be *perpendicular* to a line to the base of the flagpole. Problem 3 addresses this case.

Problem 3: Suppose that the flagpole is located on the other side of a stream. You measure the angle α of elevation of P, the top of the flagpole from a point A. Then you walk along the stream bank to a second point B at a distance b from A along a line perpendicular to the line joining A to the flagpole base Q, and measure the angle $\beta = m\angle ABQ$, as shown in Figure 14. Assume that the ground on your side of the stream is level and at the same level as the base Q of the flagpole. Find the height h of the flagpole.

Figure 14

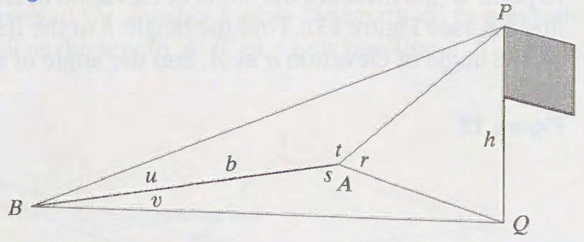


Question 3: Solve Problem 3.

Each of the Problems 1–3 involves finding the height of a flagpole by knowing the length of a baseline and the angle of inclination of the top of the flagpole from one end of that baseline. The following problem generalizes these separate cases. It asks for the height of a flagpole from the length of *any* baseline and the angles to the top and bottom of the flagpole from each end of that baseline. This problem is more challenging than the right triangle Problems 1–3. Its solution requires a multistep analysis of more than one oblique triangle, and consequently involves the Law of Sines and the Law of Cosines.

Problem 4: Suppose that there is a baseline $\overline{AB}$ of known length b, but that the baseline is not necessarily in the plane perpendicular to the flagpole at its base. The points P, Q, A, and B are all visible from one another, but do not lie in the same plane (Figure 15). Find a formula for the height h of the flagpole in terms of angles measured at the observation points A and B and the length b of the baseline.

Figure 15

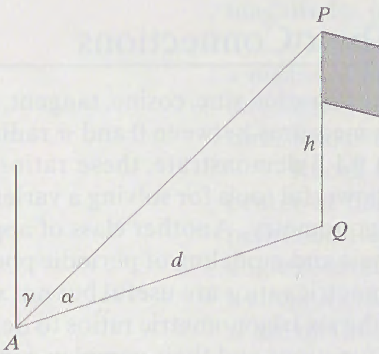


Solution For simplicity, we identify the various angles that can be measured at the observation points A and B. Let $u = m\angle ABP$, $t = m\angle PAB$, $v = m\angle ABQ$, $s = m\angle QAB$, and $r = m\angle PAQ$. Using the Law of Sines on ΔAPB and ΔAQB allows us to find AP and AQ. Then using the Law of Cosines on ΔPAQ allows us to find PQ. The details are left to you as Problem 4a.

9.1.3 Problems

1. a. Generalize Problem 1 in this section to a situation where a (vertical) flagpole is on a hill: Find the height h in terms of the distance d to the base of the pole and the angles α and γ (Figure 16), where γ is the angle between $\overrightarrow{AP}$ and the vertical.

Figure 16



b. Show that the formula you found for h in part **a** holds also for the case of a flagpole down in a valley $\left(\gamma > \frac{\pi}{2}\right)$, and in fact for any case where $\alpha + \gamma < \pi$.

c. Derive the result of Question 1 as a special case of this formula.

2. Generalize Problem 2 in this section to a situation where the base of the flagpole is on a hill above the level with the observation point A, and the line segment BAQ makes an angle ε with the horizontal.

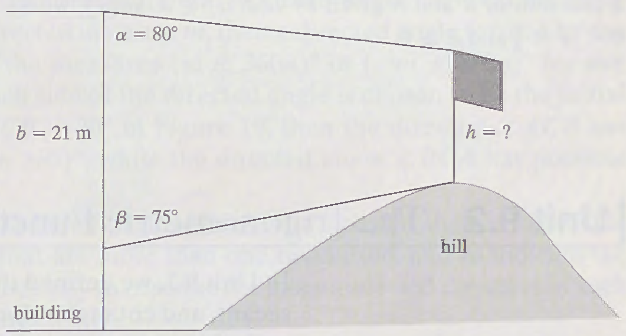
3. Generalize Problem 3 in this section to a situation where $\angle BAQ$ and $\angle AQP$ are not necessarily right angles. Specifically, assume that $\beta = m\angle ABQ$, $\gamma = m\angle BAQ$, and $\varepsilon = m\angle AQP$ are known, as well as the length $b = AB$.

4. a. Complete the derivation of the formula for the height of the flagpole in Problem 4 in this section.

b. The measure $w = m\angle PBQ$ is not needed to solve Problem 4, even though this angle could be measured from point B. (Note that this angle is *not* in general the sum of angles $u = m\angle ABP$ and $v = m\angle ABQ$, since A might not lie in the plane of triangle PBQ.) Show that w is in fact determined from the values r, s, t, u, v, and b.

5. Suppose we want to know the height h of a flagpole on a hill that we can see from a building. We measure the angle of elevation to the bottom from a low floor and an angle of depression to the top from a higher floor of the building, and measure the baseline distance b between these two floors, as shown in Figure 17. Is this information sufficient to determine the height of the flagpole or how far away it is? If it is, find these. If not, why not?

Figure 17



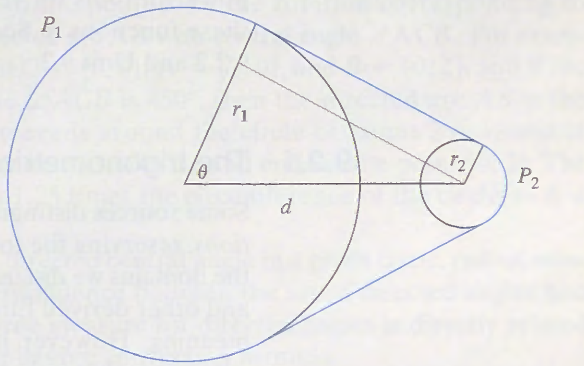
6. Consider the triangle with sides 13, 14, and 15.

a. Find the measures of the three angles in the triangle using the Law of Cosines.

b. Find the sine of the largest angle in this triangle and use the Law of Sines to find the measures of the other angles.

7. **Belt problems.** Suppose that a belt is stretched tightly over two pulleys P_1 and P_2 of radii r_1 and r_2 and whose centers are d units apart, as in Figure 18.

Figure 18



Find a formula for the total length L of the belt in terms of d, r_1, and r_2 in each case.

a. $d = 6$, and $r_1 = r_2 = 3$

b. $r_1 = r_2$ and $d = r_1 + r_2$

c. $r_1 = 12$, $r_2 = 3$, and $d = 18$

d. $r_1 > r_2$, and $d > r_1 + r_2$

8. **A Law of Cosines for Quadrilaterals:** If we know two sides and the included angle of a triangle, then SAS triangle congruence tells us that the triangle is fully determined and the Law of Cosines tells us how to find the third side.

a. Prove that if we know *three* sides and the *two* included angles of a plane quadrilateral, then the quadrilateral is fully determined, and hence that there is a SASAS quadrilateral congruence theorem.

b. Find a formula for the fourth side of a quadrilateral in terms of the other sides a, b, and c and the included angles A and B.

ANSWERS TO QUESTIONS:

1. For the particular case given, $\tan(20^\circ) = \frac{h}{350\text{ ft}} \Rightarrow h = 350\tan(20^\circ) \approx 127.40$ ft. For the general case, the height h is a function of d and α given by $h(d,\alpha) = d\tan(\alpha)$, where $0 < \alpha < \frac{\pi}{2}$ and $d > 0$.

2. Let $d = AQ$. Then $\frac{h}{d} = \tan\alpha$ and $\frac{h}{d+m} = \tan\beta$, from which $h = \frac{m\tan\alpha\tan\beta}{\tan\alpha - \tan\beta}$.

3. Let $d = AQ$. Then, from the diagram, $d = b\tan\beta$ and $h = d\tan\alpha$, so $h = b\tan\alpha\tan\beta$.

Unit 9.2 The Trigonometric Functions and Their Connections

In Unit 9.1, we defined the six trigonometric ratios, sine, cosine, tangent, cosecant, secant, and cotangent for all angles with measures between 0 and π radians (or 0° and 180°). As the problems of Section 9.1.3 demonstrate, these ratios together with the Laws of Sines and Cosines are powerful tools for solving a variety of indirect measurement problems in plane trigonometry. Another class of applications of plane trigonometry involves the analysis and modeling of periodic phenomena. For this class of applications, the trigonometric ratios are useful but not sufficient. These problems require that we extend the six trigonometric ratios to define functions for all real arguments. These real functions and their complex counterparts have a variety of remarkable algebraic, geometric, and analytic properties that are powerful tools for representing and analyzing periodic phenomena. We define these functions in Section 9.2.1 and discuss some of their applications in Section 9.2.2 and Unit 9.3.

9.2.1 The trigonometric functions

Some sources distinguish between the *trigonometric functions* and the *circular functions*, reserving the former term for functions of angles (or of their measures) with the domains we discussed in Unit 9.1. The **circular functions** refer to the sine, cosine, and other derived functions defined over all the real numbers for which they have meaning. However, in advanced mathematics, these (circular) functions and their extensions that include complex number domains are called the **trigonometric functions** and that is the wording we use here.

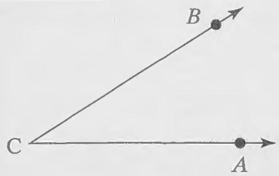
Angles, directed angles, and their measures

Recall that an *angle* in a plane is defined as the union of two rays, its *sides*, with a common initial point, its *vertex*.

In contrast, a **directed angle** is an *ordered pair* of rays with the same endpoint (its **vertex**). One side of the directed angle is prescribed as the **initial side**, the other side as the **terminal side**. The angle notation $\angle ACB$ is also used for directed angles, sometimes with the understanding that ray $\overrightarrow{CA}$ is the initial side and $\overrightarrow{CB}$ is the terminal side. Some authors use a special symbol (e.g., ∡) to distinguish angles from directed angles, but we do not.

Angles and directed angles differ significantly in the measures allowed for them. We have discussed angle measure in some detail in Section 8.2.5, so we turn now to directed angles. Whereas an (undirected) angle is usually considered to have a unique measure (and has two measures only when reflex angles are being

Figure 19



considered), a directed angle has infinitely many measures. These measures are the possible magnitudes of the rotations mapping the initial side of the angle onto its terminal side. If the rotation is counterclockwise, the directed angle is **positive**. If the rotation is clockwise, the directed angle is **negative**, and if the initial and terminal sides coincide, then the directed angle is a **zero angle**. Thus if an angle formed by two rays has undirected measure m, then a directed angle formed by the same rays may have any of the measures $(m \pm 360n)^\circ$ or $(-m \pm 360n)^\circ$ for any integer n, depending on which side of the directed angle is chosen to be the initial side. For example, if $m\angle ACB = 30^\circ$ in Figure 19, then the directed $\angle ACB$ has possible magnitude of $(30 \pm 360)^\circ$, while the directed angle $\angle BCA$ has possible magnitudes of $(-30 \pm 360)^\circ$.

The motivation behind the multiple measures for a directed angle is to provide a means of indicating turns that are more than one revolution, and to indicate the direction of such turns. Another way to describe the magnitude and direction of such turns is to consider *directed arcs*.

Recall that a *central angle* of a circle K is an angle whose vertex is at the center of that circle. If $\angle ACB$ is a central angle of circle K, and A and B are points on the circle, then A, B, and the points on arc $\widehat{AB}$ in the interior of the angle constitute the *arc subtended by* $\angle ACB$. Similarly, a **directed central angle** of a circle K is a directed $\angle ACB$ whose vertex is at the center of that circle. If A and B are on the circle, we think of the circle not only as containing the path of a point from A to B as the initial side of $\angle ACB$ is rotated to its terminal side, but we allow the path to wind around the circle more than once. This path on K traced by the point A as it is rotated to the point B by the rotation corresponding to directed $\angle ACB$ is called the **directed arc** $\widehat{\mathbf{AB}}$ **with central angle** $\angle$**ACB**. For example, in the coordinate plane $\mathbf{R}^2$, if $C = (0, 0)$, $A = (2, 0)$, and $B = (0, 2)$, and if the magnitude of the directed angle $\angle ACB$ is 450°, then the directed arc $\widehat{AB}$ is the path that starts at $(2, 0)$ and proceeds around the circle of radius 2 centered at the origin for 1.25 revolutions counterclockwise and ends at the point $(0, 2)$. The length of the directed arc $\widehat{AB}$ is 1.25 times the circumference of the circle, so it is $1.25 \cdot 2\pi \cdot 2 = 5\pi$.

For a given initial side of a directed central angle in a given circle, radian measure sets up a one-to-one correspondence between the set of directed angles and the set $\mathbf{R}$ of real numbers. Degree measure for directed angles is directly related to radian measure by the radian-degree conversion formula

$$\text{measure of directed angle in radians} = \tfrac{\pi}{180} \cdot (\text{measure of directed angle in degrees}),$$

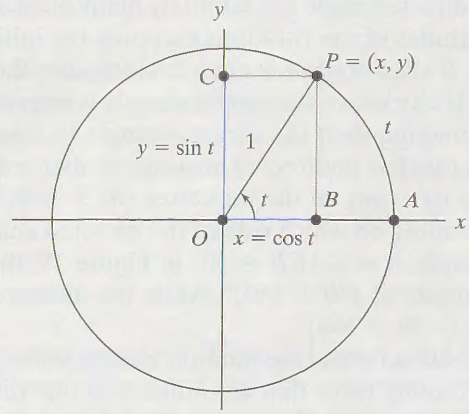
because 2π radians $= 360^\circ$.

The unit circle and wrapping function definitions of the sine and cosine functions

We begin our discussion of the trigonometric functions by recalling two ways in which they are often defined—with the unit circle and the wrapping function.

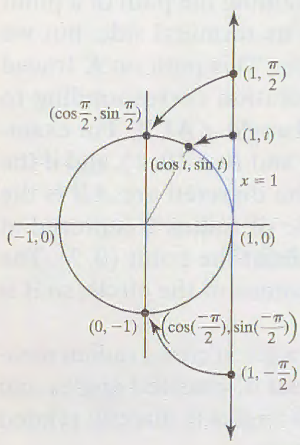
Let t be a real number, and let $\angle AOP$ be the directed angle of radian measure t with its vertex at the origin in the xy-plane, its initial side on the positive x-axis, and its terminal side intersecting the unit circle $x^2 + y^2 = 1$ at the point $P = (x, y)$. Then we *define* $\cos t = x$ and $\sin t = y$, as shown in Figure 20. Another way of conceptualizing this definition is to realize that the directed angle is itself measured by the magnitude of a rotation mapping the initial side onto its terminal side. So an equivalent definition is to define $\cos t$ and $\sin t$ to be the first and second coordinates of the image of $(1, 0)$ under a rotation of t radians.

Figure 20



A major advantage of using the unit circle is that, when $-\pi < m < AOP \leq \pi$, due to Theorem 9.2, $m\angle AOP$ in radians is numerically equal to the length of the arc of the circle subtended by $\angle AOP$. This way of defining the trigonometric functions is usually referred to as the **unit circle definition**.

Figure 21



In the late 1940s, a **wrapping function definition** of the trigonometric functions was proposed. It is also based on the unit circle but in an interestingly different way.

The wrapping function definition can be described dynamically as follows: Begin with a unit circle and the line $x = 1$, as shown in Figure 21. Imagine wrapping this line around the circle in two directions. The ray above the x-axis is wrapped counterclockwise around the circle. Many points on the line map onto the same point of the circle. To find the image of a point on the line, we match the distance of this point above the x-axis with the length of an arc of the unit circle as measured from counterclockwise $(1, 0)$. For instance, the point $\left(1, \frac{\pi}{2}\right)$ on the line is $\frac{\pi}{2}$ above the x-axis. Since the circumference of the unit circle is 2π, $\frac{\pi}{2}$ is a quarter of the circumference. So its image will be a quarter-way around the circle, at the point $(0, 1)$. In analogous fashion, the image of $\left(1, \frac{\pi}{4}\right)$ on the line is the point $\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$. The ray below the x-axis is mapped clockwise around the circle. So the image of $\left(1, -\frac{\pi}{2}\right)$ on the line is the point $(0, -1)$ on the circle. In general, the images of all points of the form $(1, k + 2n\pi)$, where n is an integer, map onto the same point of the circle.

With this conception, $\cos x$ and $\sin x$ are defined as the first and second coordinates of the image of $(1, x)$. This definition is far removed from triangles, but it has some advantages. It gets us straight to thinking in terms of radians rather than degrees. The periodicity of the sine and cosine functions is an immediate consequence (it also is an immediate consequence of the definition in terms of arcs and rotations). And, by identifying the real number x with the length x, the trigonometric functions are seen as functions of real numbers, enabling them to be treated as real functions.

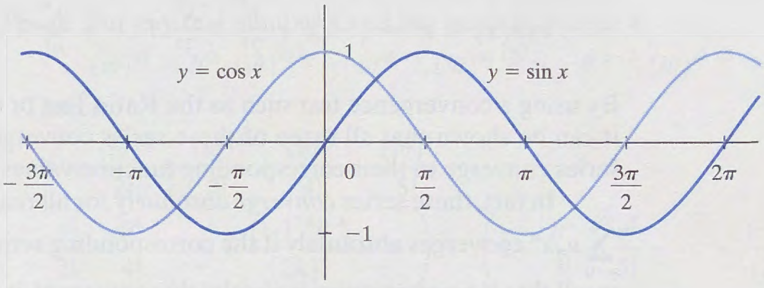
There are, however, a number of pedagogical disadvantages of the wrapping function approach. The lack of connection with triangle trigonometry leads some students to think these are different sines and cosines than those they have seen before. The lack of experience with arc length makes it difficult for some students to identify values of the functions. And the lack of experience with the idea of wrapping makes the definition difficult to apply.

Both the unit circle and wrapping function definitions clearly establish the sine and cosine functions as real functions that are periodic with a period of 2π. By way of contrast, the sine and cosine trigonometric ratios are defined only for acute and obtuse angles.

The unit circle and wrapping function definitions of the sine and cosine for $0 < t < \frac{\pi}{2}$ and $\frac{\pi}{2} < t < \pi$ can be shown to be consistent with the ratio definitions of these functions, as follows: Refer to Figure 20. There $t = m\angle POA$. If $0 < t < \frac{\pi}{2}$, then from the ratio definitions $\sin t = \frac{PB}{OP} = PB = y$ and $\cos t = \frac{OB}{OP} = OB = x$. For obtuse angles, $\sin t$ and $\cos t$ are defined as the ratios $\frac{y}{r}$ and $\frac{x}{r}$, respectively (see Section 9.1.1). Here $r = 1$, so again we have $\sin t = y$ and $\cos t = x$ as in both the unit circle and wrapping function definitions.

The familiar graphs (Figure 22) of the sine and cosine functions—*sine waves* of infinite extent—are only possible when the domain of these functions is the set **R**.

Figure 22



Once $\cos t$ and $\sin t$ have been defined for all real numbers t, the other four trigonometric functions are defined either as their reciprocals or their quotients.

$$\tan t = \frac{\sin t}{\cos t}, \quad \cot t = \frac{\cos t}{\sin t}, \quad \sec t = \frac{1}{\cos t}, \quad \text{and} \quad \csc t = \frac{1}{\sin t},$$

for all real numbers t for which the denominators are not zero. As an immediate consequence of being the reciprocals of the sine and cosine functions, the secant and cosecant functions are periodic of period 2π. The tangent and cotangent functions are also periodic, but their period is π (see Problem 1).

Obtaining values of the trigonometric functions

We could obtain the values of trigonometric functions from the values of the trigonometric ratios in the interval $\left[0, \frac{\pi}{2}\right]$. But in practice, this is difficult to do. We would have to obtain certain values of the functions $\left(\text{those for } \frac{\pi}{2}, \frac{\pi}{3}, \frac{\pi}{4}, \text{and } \frac{\pi}{6}\right)$ from theorems of geometry, employ a variety of trigonometric identities to obtain values for the sums, differences, and multiples of these numbers, and interpolate between values to determine other values to the desired accuracy. This is how Ptolemy constructed his tables of lengths of chords, which preceded later tables of sines. Problem 6 in Section 9.3.1 illustrates how trigonometric identities can be used to obtain exact values of $\sin \frac{\pi}{2^k}$ and $\cos \frac{\pi}{2^k}$ for any integer $k \geq 2$, values which are critical in this process.

An easier way to obtain all values to any desired accuracy is by the representation of these functions as power series by using the Taylor series expansion. Recall from calculus that, for a given real function f with derivatives of all orders on its domain, the Taylor series centered at point a in the domain of f is

$$f(x) \approx \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n$$

$$= f(a) + f^{(1)}(a)(x-a) + \frac{f^{(2)}(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \cdots.$$

When $a = 0$, the Taylor series is called the **Maclaurin series for f**.

From calculus, the Maclaurin series expansion of many functions can be computed rather easily. Principal among these are the Maclaurin series for the cosine and sine functions and for the exponential function defined by $\exp(x) = e^x$.

$$\cos(x) = 1 - \frac{x^2}{2!} + \cdots + (-1)^n \frac{x^{2n}}{(2n)!} + \cdots = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!}$$

$$\sin(x) = x - \frac{x^3}{3!} + \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} + \cdots = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!}$$

$$\exp(x) = e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

By using a convergence test such as the Ratio Test or some other convergence test, it can be shown that all three of these series converge for all real x. In fact, these series converge to their corresponding function values for all real x.

In fact, these series *converge absolutely* for all real x. Recall that a power series $\sum_{n=0}^{\infty} a_n x^n$ converges absolutely if the corresponding series $\sum_{n=0}^{\infty} |a_n||x|^n$ converges. Also recall that if a power series is absolutely convergent, it is necessarily convergent, but that the converse is not true in general. For example, the power series $\sum_{n=1}^{\infty} \frac{(-1)^n}{n} x^n$ converges when $x = 1$ (it becomes an alternating series with terms decreasing to 0) but it does not converge absolutely because the series $\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} + \cdots$, the harmonic series, diverges).

Extending the domain of the trigonometric functions to include complex numbers

It is unfortunate that nearly all modern calculus texts fail to include any discussion of complex numbers. There are many fruitful connections between real and complex functions that are relatively easy to develop. Here is one important example. Suppose that we replace the real variable x in the power series representations of $\sin x$, $\cos x$ and e^x by a complex variable z. Then, because $|z|$ is a real number for any complex number z, and because each of these series converges absolutely for all real numbers, the power series for $\sin z$, $\cos z$ and e^z converge absolutely for all complex z. Just as for real series, absolute convergence implies convergence, so the corresponding complex power series formulas can be regarded as *definitions* of $\sin z$, $\cos z$ and e^z for all complex numbers z,

$$\cos(z) = 1 - \frac{z^2}{2!} + \cdots + (-1)^n \frac{z^{2n}}{(2n)!} + \cdots = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!},$$

$$\sin(z) = z - \frac{z^3}{3!} + \cdots + (-1)^n \frac{z^{2n+1}}{(2n+1)!} + \cdots = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{(2n+1)!},$$

$$\exp(z) = e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots + \frac{z^n}{n!} + \cdots = \sum_{n=0}^{\infty} \frac{z^n}{n!}.$$

These are complex extensions of the real cosine, sine, and exponential functions. That is, their values agree with the values of these real functions when z is a real number. As complex functions, the sine and cosine functions are still periodic with period 2π, and interestingly enough, the exponential function is periodic with period $2\pi i$ (see Problem 6).

Euler's Formula

One of the most remarkable consequences of the extension of the sine, cosine, and exponential functions to the complex numbers is the following result, which relates trigonometric and exponential functions.

Theorem 9.7 **(Euler's Formula):** For any real number θ,

$$e^{i\theta} = \cos\theta + i\sin\theta.$$

Proof: For any real number θ and any positive integer k,

$$(i\theta)^{4k} = \theta^{4k},\ (i\theta)^{4k+1} = i\theta^{4k+1},\ (i\theta)^{4k+2} = -\theta^{4k+2},\ (i\theta)^{4k+3} = -i\theta^{4k+3}.$$

Therefore,

$$e^{i\theta} = 1 + i\theta + \frac{(i\theta)^2}{2!} + \frac{(i\theta)^3}{3!} + \frac{(i\theta)^4}{4!} + \frac{(i\theta)^5}{5!} + \cdots$$

$$= \left(1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \cdots + \frac{(-1)^k\theta^{2k}}{(2k)!} + \cdots\right) + i\left(\theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \cdots + \frac{(-1)^k\theta^{2k+1}}{(2k+1)!} + \cdots\right)$$

$$= \cos\theta + i\sin\theta$$

The most famous consequence of Euler's Formula relates the important mathematical numbers e, π, i, 1, and 0.

Corollary: $e^{i\pi} = -1$, or, equivalently, $e^{i\pi} + 1 = 0$.

9.2.1 Problems

1. Explain why the tangent and cotangent functions have period π while the sine and cosine functions have period 2π.

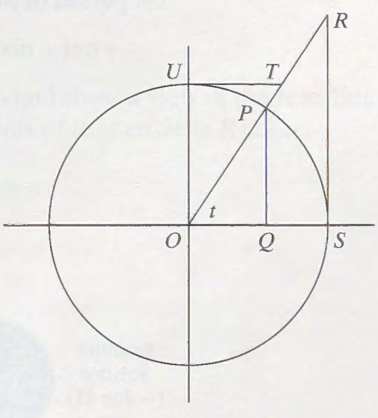
2. The identities in this problem are known as the Pythagorean identities.
   a. Explain why $\sin^2 A + \cos^2 A = 1$ for any acute or obtuse angle A using the trigonometric ratio definition of sine and cosine.
   b. Explain why $\sin^2 x + \cos^2 x = 1$ for any x using the unit circle definitions of the sine and cosine functions.
   c. Divide both sides of the identity in part **b** by appropriate expressions to obtain two more identities similar to that one.

3. The identities $\sin(-x) = -\sin x$ and $\cos(-x) = \cos x$ are not meaningful for the trigonometric ratio definition of sine and cosine. Derive these identities from one of the other definitions of sine and cosine.

4. a. In the wrapping function, name two points on the line $x = 1$ whose image is $\left(-\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$.
   b. What specific values of sines and cosines are obtained from the information in part **a**?

5. Figure 23 suggests how four of the trigonometric functions got their names. O is a circle of radius 1 and t is the measure of the acute angle POQ. Each of the six trigonometric functions of t has length equal to the length of a segment on the figure. Find the segment for each function.

Figure 23



6. Prove that the exponential function defined on C, $\exp(z)$, is periodic with period $2\pi i$.

7. Prove the corollary to Theorem 9.7.

8. By applying Theorem 9.7 to the exponential $e^{i(\theta+\phi)}$, derive the familiar formulas for $\sin(\theta + \phi)$ and $\cos(\theta + \phi)$.

9. Use Theorem 9.7 to deduce formulas for $\sin x$ and $\cos x$ in terms of e and x. (*Hint*: Replace θ by $-\theta$ to deduce a corollary to the theorem.)

9.2.2 Modeling with trigonometric functions

Many phenomena, including sound, electromagnetic and water waves, mechanical vibrations, seasonal, biological and economic cycles, tides, and the motion of celestial bodies and orbiting spacecraft, are driven by motions that are periodic or approximately periodic. Manufactured objects such as electric generators and motors, internal combustion engines, and axles of cars and trucks all rotate while in use, so are described by periodic functions. The rotation of an alternating current generator produces a current with voltage, or AC. It is the reason why the complex exponential function (Problem 6 of Section 9.2.1) is so important in electrical engineering. The trigonometric functions are perhaps the most familiar examples of periodic functions.

Modeling periodic phenomena

Recall that a function f is **periodic** if and only if there exists a positive number p such that $f(x + p) = f(x)$ for all x. The smallest such number p (if there is a smallest number) is the **period** of the function.

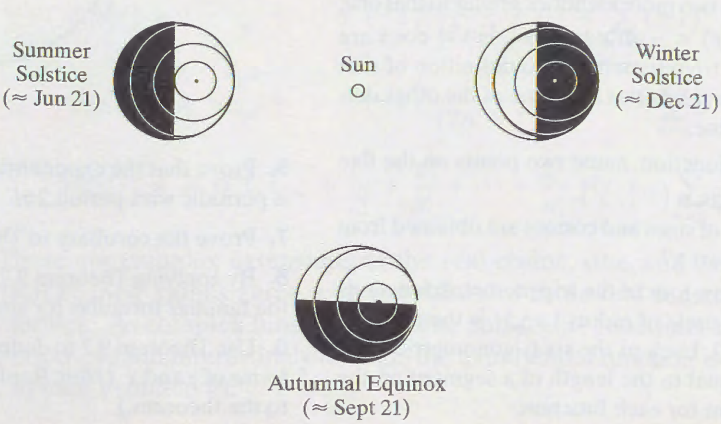
From the definitions of $\cos x$ and $\sin x$ in terms of rotations, $\cos(x + 2\pi) = \cos x$ for all x. Furthermore, the cosine function is decreasing on the interval $[0, \pi]$ and increasing on the interval $[\pi, 2\pi]$, so there is no number p less than 2π for which $\cos(x + p) = \cos x$ for all x. This implies that 2π is the period of the cosine function.

Question: Show that if $c > 0$, the function f defined by $f(x) = \sin(cx)$ for all real numbers x is periodic with period $\frac{2\pi}{c}$.

The length of daylight

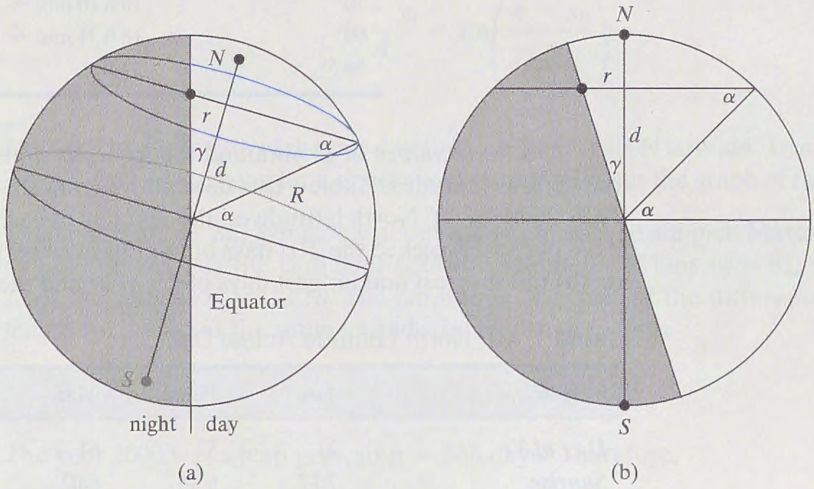
The following example illustrates how trigonometric functions can model periodic phenomena. The earth tilts on its axis at a fairly constant angle of $\gamma \approx 23.5°$ to the perpendicular to the plane of its orbit around the sun. As a result of this tilt, the length of the daily period of daylight in the northern hemisphere varies throughout the year from the longest period of daylight near the summer solstice (≈June 21) to the shortest period of daylight near the winter solstice (≈December 21) (see Figure 24).

Figure 24



If the north pole N is the upper end point of the earth's axis, then Figure 25a describes the position of Earth at the summer solstice. Imagine the Sun at the right (over 11,600 Earth diameters away!), so the shaded part is night. The equator and the circle for latitude α are shown. From this diagram, we see that the period of daylight on the longest day of the year varies with the latitude. For instance, at the equator half the circle of latitude is in daylight, so half the day is in daylight. At the latitude α drawn, about $\frac{2}{3}$ of the day would be in daylight. Figure 25b shows a cross-section viewed from the plane of the Equator; the Sun is then 23.5° above the horizontal.

Figure 25



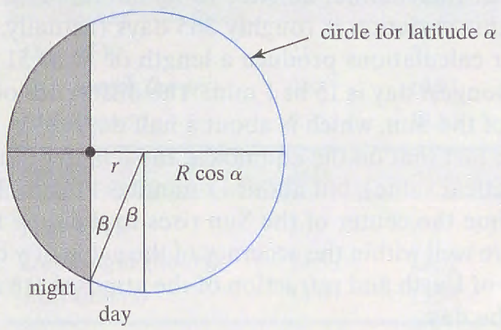
EXAMPLE 1 Find the length of the longest day as a function of the latitude α in degrees.

Solution We need only right-triangle trigonometry. Refer to Figure 25. Let d be the distance from the plane of the circle of latitude α to the center of Earth. Then $\sin \alpha = \frac{d}{R}$, and also $\tan \gamma = \frac{r}{d}$. Solving each equation for d and equating the solutions, we find that the distance r and Earth's radius R are related by

$$r = R \sin \alpha \tan \gamma.$$

If we slice through the earth at latitude α and show a view of the resulting circle from Figure 25, we obtain Figure 26. The radius of that circle is $R \cos \alpha$.

Figure 26



In Figure 26, $\sin \beta = \frac{r}{R \cos \alpha} = \tan \alpha \tan \gamma$, from which $\beta = \sin^{-1}(\tan \alpha \tan \gamma)$. (Notice that R, the earth's radius, cancels out.)

The portion of the day that is daylight at this latitude equals the portion of the arc of the circle that is blue. From Figure 26, that portion is $\frac{2\beta + 180°}{360°}$. Since there are 24 hours in a day, the length D of the longest day in hours is given by $D = 24\frac{2\beta + 180°}{360°}$.

Since $\gamma \approx 23.5°$, $\tan \gamma \approx \tan 23.5° \approx 0.435$, and $\beta \approx \sin^{-1}(0.435 \tan \alpha)$.

Evaluating D at various latitudes α, we have the following:

Latitude°	Hours in Longest Day
0°	12 h = 12.00 h
10°	12 h 35 min ≈ 12.58 h
20°	13 h 12 min = 13.20 h
30°	13 h 56 min ≈ 13.93 h
40°	14 h 51 min ≈ 14.85 h
50°	16 h 10 min ≈ 16.17 h
60°	18 h 31 min ≈ 18.52 h
66.5°	24 h

Do these values of D obtained theoretically agree with actual longest days at the various latitudes? Table 1 lists data drawn from the *World Almanac and Book of Facts 2000* for 40° North latitude on the 21st day of each month of the year 2000. We have purposely picked the 21st days of months because they provide (within a minute or so) the shortest and longest days of the year and the solstices.

Table 1 40° North Latitude Actual Data

Month	Jan.	Feb.	Mar.	April	May	June
Day of Yr	21	52	81	112	142	173
Sunrise	7:17	6:46	6:02	5:13	4:40	4:31
Sunset	17:05	17:42	18:13	18:35	19:14	19:32
Day length (hours)	9.83	10.93	12.18	13.53	14.57	15.02

Month	July	Aug.	Sept.	Oct.	Nov.	Dec.
Day of Yr	203	234	265	295	326	356
Sunrise	4:49	5:17	5:47	6:17	6:52	7:18
Sunset	19:24	18:48	17:58	17:12	16:39	1638
Day length (hours)	14.58	13.52	12.18	10.92	9.78	9.33

At a given latitude α, the length of the daily period of daylight varies from its maximum at the summer solstice to its minimum at the winter solstice and the period of this variation is roughly 365 days (actually, 365.2425 days—see Section 6.1.2). Our calculations produce a length of 14 hr 51 min for the longest day, but the actual longest day is 15 hr 1 min. The difference of 10 minutes is due mainly to the width of the Sun, which is about a half degree, or 1/720 of a circle. This is verified by the fact that on the equinoxes, the length of the day is not exactly 12 hours (the theoretical value), but about 11 minutes longer. If we judge the length of day from the time the center of the Sun rises to the time the center sets, then our calculations are well within the accuracy of the value of γ used for the tilt of Earth. The oblateness of Earth and refraction of the atmosphere at the horizon also affect the length of the day.

The data in the table are periodic. If you consult an almanac for the year in which you read this page, you will find values for the days of the year that are very close to the ones given. Furthermore, graphing the values suggests that a sine wave will fit the data reasonably well.

Sine waves are images of the graph of $y = \sin x$ under stretches and translations. By the Graph Translation Theorem (Theorem 7.21), a translation image of the wave that is rotation symmetric about (x_0, y_0) rather than $(0, 0)$ has equation

$y - y_0 = \sin(x - x_0)$. From the Question of this section, if we wish the period of the sine wave to be p rather than 2π, and still maintain that symmetry, we may write

$$y - y_0 = \sin\left(\frac{x - x_0}{\frac{p}{2\pi}}\right).$$

Finally, if we wish the amplitude of the sine wave to be A rather than 1, we can use the equation

$$\frac{y - y_0}{A} = \sin\left(\frac{x - x_0}{\frac{p}{2\pi}}\right).$$

EXAMPLE 2 Let $L(d)$ be the length of the day d of the year 2000 at 40°N latitude. Using the data in Table 1, find an equation for a sine wave that approximates the graph of the function L.

Solution The length of day has symmetry about the equinoxes, so we pick March 21, the 81st day of the year, as the center of rotation symmetry. Then $x_0 = 81$. On this day, $L(d) = 12.18$, so $y_0 = 12.18$. The amplitude A is half of the difference of the day length (in hours) at the summer and winter solstices. Thus

$$A = \frac{1}{2}(15.02 - 9.33) = 2.85.$$

The year 2000 was a leap year, so $p = 366$ days. Therefore,

$$\frac{L(d) - 12.18}{2.85} = \sin\left(\frac{d - 81}{\frac{366}{2\pi}}\right).$$

Consequently, $L(d) = 12.18 + 2.85\sin\left[\frac{2\pi}{366}(d - 81)\right]$.

This model produces day length data for the 21st of each month in the year 2000 (Table 2) that are close to but not identical to the data in Table 1.

Table 2 40° North Latitude Model Values

Month	Jan.	Feb.	Mar.	April	May	June
Day of Yr	21	52	81	112	142	173
Day length (hours)	9.83	10.93	12.18	13.53	14.57	15.02
L(d) (hours)	9.75	10.82	12.18	13.62	14.63	15.02
Month	**July**	**Aug.**	**Sept.**	**Oct.**	**Nov.**	**Dec.**
Day of Yr	203	234	265	295	326	356
Day length (hours)	14.58	13.52	12.18	10.92	9.78	9.33
L(d) (hours)	14.63	13.58	12.13	10.74	9.70	9.34

Describing plane curves with angles as parameters

Many important plane curves can be described most conveniently and completely by parametric equations in which the parameter is a suitably selected directed angle. In such cases, the tools of trigonometry often play an important role in the development of the description.

Theorem 9.8(a) The circle of radius r centered at the point (h, k) has parametric equations

$$x = h + r\cos\theta, \qquad y = k + r\sin\theta,$$

where the parameter θ is the directed angle with vertex at (h, k) with initial side parallel to the positive x-axis and terminal side joining (h, k) to the point (x, y). As θ increases from 0 to 2π, the corresponding point (x, y) traces out the circle with the rectangular equation $(x - h)^2 + (y - k)^2 = r^2$ once in the counterclockwise direction starting from the point $(h + r, k)$.

Proof: In Problem 5, you are asked to show that any point satisfying the parametric equations is on the circle with the indicated rectangular equation, and vice versa. ┘

By stretching the circle of Theorem 9.8(a), we obtain a more general theorem involving ellipses.

Theorem 9.8(b) The ellipse centered at (h, k) with semimajor axis of length a parallel to the x-axis, and semiminor axis of length b parallel to the y-axis, has parametric equations

$$x = h + a\cos\theta, \qquad y = k + b\sin\theta,$$

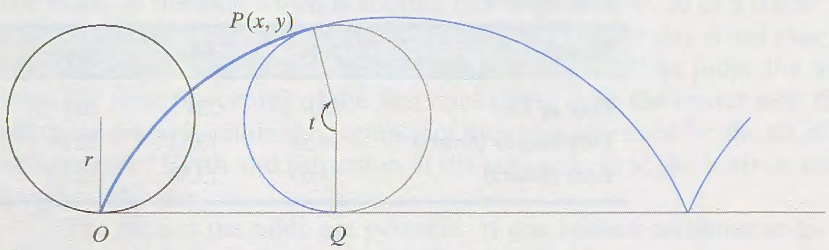
with θ defined as in Theorem 9.8a. As θ increases from 0 to 2π, (x, y) traces out the ellipse with the rectangular equation $\frac{(x-h)^2}{a^2} + \frac{(y-k)^2}{b^2} = 1$ once in the counterclockwise direction starting from the point $(h + a, k)$.

Proof: A proof is again left to you (see Problem 6). ┘

The next problem is concerned with finding parametric equations involving an angular parameter for a famous plane curve called a *cycloid*. The solution of this problem makes significant use of concepts and methods from trigonometry.

EXAMPLE 3 A wheel of radius r rolls along a straight level track. The point P on the rim of the wheel that is in contact with the track when the wheel began to roll describes a plane curve C as the wheel rolls. If t is the central angle in radians through which the wheel has rolled to place the wheel in its current position $P(x, y)$, find equations for x and y in terms of the parameter t.

Figure 27



Solution: In Figure 27, the length of $\overline{OQ}$ is equal to the length of the circular arc $\widehat{QP}$ because both represent the distance that the point P has traveled as the central angle t increased from 0 to t. Therefore, because the radius of the wheel is r, the x-coordinate of Q is tr if t is measured in radians. Therefore, the x-coordinate of the point P is $tr - r\sin t$ and the y-coordinate of the point P is $r - r\cos t$. (Note that $\cos t < 0$ for the position of P in Figure 27.) Thus, the cycloid is described by the parametric equations

$$\begin{cases} x = r(t - \sin t) \\ y = r(1 - \cos t) \end{cases}.$$

We could cite many other types of problems for which plane trigonometry is useful. The samples that we have selected in this section and in Section 9.1.3 give some indication of the diversity of these problems as well as the tools and concepts of trigonometry that are used for their solution. In the applications that we have considered, there are other trigonometric concepts and tools that have not been used such as the addition formulas, the formulas for rotation of axes, polar coordinates, and the trigonometric form of complex numbers. Their widespread use attests to the fundamental importance of trigonometry.

9.2.2 Problems

1. Give the amplitude and period of the function f defined by $f(x) = 3\sin(4x) + 5$.

2. Give a center of rotation symmetry for the function g with $g(x) = A\sin(B(x + C)) + D$.

3. Can a sine wave always be described by an equation involving the cosine as the only trigonometric function? Why or why not?

4. Using information from Example 1, find an equation for $L(d)$, the length of day d of the year 2000 on Earth at latitude 30°S.

5. Prove Theorem 9.8a.

6. Prove Theorem 9.8b.

7. a. Explain why $\begin{cases} x = a\sec\phi + h \\ y = b\tan\phi + k \end{cases}$ are parametric equations for the hyperbola with the rectangular equation $\frac{(x-h)^2}{a^2} - \frac{(y-k)^2}{b^2} = 1$.

   b. Let $a = b = 1$ and $h = k = 0$. Graph this hyperbola using a function grapher. Explain why the order of tracing out the hyperbola is as it is.

8. Many periodic functions are not trigonometric functions. For example, the **square wave function** s is defined by

$$s(x) = \begin{cases} 1 \text{ if } & 2k\pi < x \le (2k+1)\pi \\ -1 \text{ if } & (2k+1)\pi < x \le 2(k+1)\pi \end{cases} \text{ for any integer } k.$$

   a. Sketch the graph of the square wave function on the interval $[-3\pi, 3\pi]$. Then use a graphing calculator or computer to plot the same function on the same interval. Explain any differences that you see between the resulting graph and the graph that you sketched.

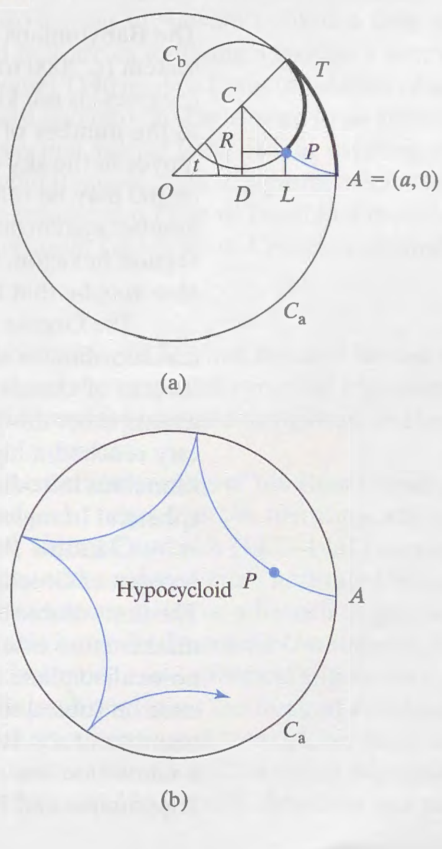
   b. Use a graphing calculator or computer to graph the function

$$p(x) = \frac{4}{\pi}\left[\sin(x) + \frac{1}{3}\sin(3x) + \frac{1}{5}\sin(5x) + \frac{1}{7}\sin(7x)\right]$$

   together with the square wave function on the interval $[-3\pi, 3\pi]$. Discuss the closeness of the two graphs.

9. Suppose that a circle C_a of radius a is fixed with center at the origin O and that a circle C_b inside of C_a with radius $b < a$ and tangent to the circle C_a at the point $A = (a, 0)$ begins to roll in a counterclockwise direction around the inside of C_a (Figure 28a). Then the point $P = (x, y)$ on the circle that was initially at A describes a curve inside of the circle C_a called a **hypocycloid** (Figure 28b). We wish to find an equation for the hypocycloid in terms of t, where $t = m\angle AOC$ (in radians).

Figure 28



Let C be the (moving) center of circle C_b, let T be the point of tangency of the two circles, and let $s = m\angle PCT$.

a. Explain why $s = \frac{at}{b}$ and $s - t = \frac{a-b}{b}t$.

b. Show that parametric equations for the hypocycloid are given by

$$\begin{cases} x = (a-b)\cos(t) + b\cos\left(\frac{a-b}{b}t\right) \\ y = (a-b)\sin(t) + b\sin\left(\frac{a-b}{b}t\right) \end{cases}.$$

c. If $b = \frac{a}{4}$, show that these parametric equations for the hypocycloid simplify to

$$\begin{cases} x = a\cos^3(t) \\ y = a\sin^3(t) \end{cases}$$

and that a rectangular equation for this hypocycloid is

$$x^{2/3} + y^{2/3} = a^{2/3}.$$

(*Hint*: Use identities for $\sin(3t)$ in terms of $\sin t$ and $\cos(3t)$ in terms of $\cos t$.)

10. If the circle C_b in Problem 9 rolls along the *outside* of the circle C_a, a point P on the circumference of C_b describes a curve in the plane called an **epicycloid**. Assuming that the point $P = (x, y)$ is initially at the point $(a, 0)$ and that C_b rolls counterclockwise around C_a, show that

$$\begin{cases} x = (a+b)\cos(t) - b\cos\left(\frac{a+b}{b}t\right) \\ y = (a+b)\sin(t) - b\sin\left(\frac{a+b}{b}t\right) \end{cases},$$

where $t = m\angle AOC$ (in radians), as in Problem 9.

11. Use integration to prove that the area under one arch of the cycloid

$$x = r(t - \sin t) \qquad y = r(1 - \cos t)$$

is $3\pi r^2$ and that the length of this arch is $8r$.

ANSWERS TO QUESTION:

1. For all $x \in \mathbf{R}$, $f(x) = \sin(cx) = \sin(cx + 2\pi) = \sin(c(x + \frac{2\pi}{c}))$, and $\frac{2\pi}{c}$ is the smallest number p for which $\sin(cx) = \sin(c(x + p))$, so $\frac{2\pi}{c}$ is the period of f.

9.2.3 The historical and conceptual evolution of trigonometry

Babylonian and Greek contributions

The Babylonians were first to introduce degree measure and a spherical coordinate system (c. 2000 to 1600 B.C.). The rationale for the division of a circle into 360 parts (degrees) is not known for certain. It is often theorized to be 360 because 360 is close to the number of days of the year (and so one degree is close to the amount a star moves in the sky each day) and to the fact that 360 has so many factors. The choice of 360 may be related to the fact that the Babylonians used a sexagesimal (base 60) number system and a circle is divided neatly into 6 parts by the vertices of an inscribed regular hexagon, with the resulting chord lengths equal to the radius of the circle. It also may be that the base system was related to the choice of the degree.

The Greeks adopted and refined both degree measure and the work with spherical coordinates. Hipparchus of Nicaea (c. 180–125 B.C.) prepared the first tables of lengths of chords subtending given circular arcs, and used these tables to calculate, among other things, longitudes and latitudes on the celestial sphere. Greek trigonometry reached a high point with Menelaus (98 A.D.). In his seminal work *Sphaerica*, Menelaus introduced the concept of a spherical triangle and proved theorems about spherical triangles analogous to those Euclid had for plane triangles.

Claudius Ptolemy (c. 100–178 A.D) wrote a comprehensive and definitive account of Greek astronomy based on the earlier work of Hipparchus and Menelaus. The first volume of this 13-volume treatise was largely devoted to an exposition of the mathematics essential to his discussion of astronomy, primarily content from what is now called plane and spherical trigonometry. Ptolemy's books had a profound influence on subsequent developments in astronomy throughout the world until the sixteenth century. Islamic writers referred to these books as the *Almagest* (the greatest), a name that has now been generally adopted for Ptolemy's treatise. The tables of Hipparchus and Ptolemy (discussed in Section 9.1.2) were precursors of later tables of sines. Until the advent of calculators in the 1970s, tables of sines and tangents were essential for anyone who wished to apply trigonometry.

Hindu and Arab contributions

In the ninth and tenth centuries, Hindu and Arab mathematicians further refined the trigonometric concepts and methods developed in Ptolemy's *Almagest.* In particular, they introduced the modern sine concept from the Greek work on chords and arcs. The historical development is quite similar to the way that Hindu-Arabic decimal notation became the worldwide standard. Indian mathematicians defined the cosine, tangent, and cotangent (of course, not using those names). The Muslim mathematician and astronomer Mohammed ibn Jabir ibn Sinan Abu Abdullah al-Battani (850–929) developed the spherical Law of Cosines to calculate the measure of the arc in space between a planet and the Sun. It was given in his astronomical treatise whose title translates as *On the science and number of stars and their motions.* He computed tables of sines, tangents, and cotangents for angles from 0° to 90°. The Persian Abu'l-Wafa (940–998) first discovered the spherical Law of Sines, and Al Buruni (973–1048) later established the corresponding result for plane triangles. This order of discovery may seem curious to us now, but it should be remembered that a major impetus for the study of trigonometry from ancient times through the sixteenth century came from astronomy, and for astronomy, spherical trigonometry was more critical than plane trigonometry.

Almost none of these achievements of the Indian and Islamic mathematicians were known in the medieval West. Europe was struggling through the Dark Ages, and it wasn't until the 12th century that Latin translations were made of ancient mathematical treatises. The Hindu mathematician Aryabhata first used the half-chord, which he called "jyardha", later shortened to "jya". The Arabs translated this word as *jiba*, which meant "nothing" in Arabic, and which is written (as all Arabic is written) without vowels as *jb*. Later readers thought that *jb* stood for the word *jaib* (also written as *jb*), which means "inlet or bay" but also can mean "bosom". (With a little imagination, you can picture a half-chord and half-arc as outlining a mother's arm as she is cradling a child.) Robert of Chester around 1140 made a Latin translation of a treatise by al-Khwarizmi and translated the word *jaib* into the Latin word *sinus* (which means "inlet, bay, and bosom"). Some books say that *jaib* (or *jayb*) had the meaning of "chord of an arc" but that the Europeans thought it meant "fold of a garment", for which the Latin was again "sinus". Some sources believe that Plato of Tivoli first introduced the word "sinus", around 1116–1136; others credit Gherardo of Cremona, around 1150.

European contributions

Although Fibonacci, in his *Practica Geometriae* of 1220, had initiated the use of plane trigonometry, until 1450 the focus in trigonometry was on spherical trigonometry. In the late fifteenth century, because of the need for accurate navigation and surveys of land, plane trigonometry became very important.

Trigonometry was studied by most mathematicians of the Renaissance. By the sixteenth century it began to be treated as a "subject" in the literature, and acquired the status of a branch of mathematics. Georg Peurbach (1423–1461) corrected the Latin translations of Ptolemy's *Almagest* and produced more accurate trigonometric tables. The first systematic treatment of trigonometry as a branch of geometry was given by Regiomontanus (1436–76) in his book *De Triangulis Omnimodus*, in which he gave an axiomatic development of both plane and spherical trigonometry within Euclid's framework. Regiomontanus's book continues the work of Peurbach. Others in the fifteenth and sixteenth centuries continued to work on more accurate trigonometric tables, including George Joachim Rheticus (1514–1613), Nicolaus Copernicus (1473–1543), and Bartholomaus Pitiscus (1561–1613). Rheticus had published chapters of Copernicus's famous *De revolutionibus orbium Coelestium*. These dealt with planar and spherical trigonometry and served as a compendium of the trigonometry pertinent to the astronomy at that time.

The word "trigonometry" was introduced by Pitiscus in 1595 in his book *Trigonometria*. The word "goniometry" was introduced to refer to that part of trigonometry dealing only with angles. Pitiscus also introduced the formulas for the sine and cosine of the sum and difference of two angles, which up to then had been calculated on the basis of tables. But although mathematicians had the formulas, they didn't have the language of algebra to work with them.

François Viète (1540–1603) brought the algebra of trigonometry into being with several treatises he wrote in the late 16th century. He also extended the tables of Rheticus and added to the trigonometric identities that had been established by Ptolemy. In 1615 his versions of formulas for $\sin(nA)$ as functions of $\sin A$ and $\cos A$ were published posthumously.

As functions became central to algebra, many new trigonometric formulas were developed. Important relationships were established in the seventeenth and eighteenth centuries between the trigonometric functions and powers and multiples of angles, series, continued fractions, polar coordinates, and other ideas. The names associated with these developments are quite familiar. Isaac Newton (1642–1727) used polar coordinates in connections with tangents, curvature, and rectification of curves. He derived a formula for the radius of curvature using polar coordinates and trigonometric ratios, and he connected the trigonometric functions to infinite series.

Leonhard Euler (1707–1783) wrote a textbook, *Introductio in analysin infinitorum*, in 1748, which has been compared to the *Elements*, doing for analysis what Euclid did for geometry. In this textbook, Euler popularized the definition of the trigonometric functions as ratios and derived their series expansions using binomial series and a limiting argument. We have noted earlier that Euler produced the formula $e^{ix} = \cos x + i \sin x$, unifying the power function with the trigonometric functions and the imaginary number i. Euler is responsible for expanding much of the theory of trigonometry in the later part of the 18th century. He analyzed all the trigonometric functions, systematically listed all the usual formulas of goniometry, emphasized the periodicity of the trigonometric functions, etc. He connected trigonometry with coordinate geometry and calculus. He expanded the distance function in mathematical astronomy as a series involving trigonometric ratios. He looked back at the spherical trigonometry of Menelaus and showed how the various theorems could be derived algebraically one from the other.

In 1798–1799, Sylvestre-François Lacroix (1765–1843) published an influential textbook on trigonometry called *Traite elementaire de trigonometrie rectiligne et spherique et application de l'algebre a la geometrie*. This textbook went through many editions and was translated into many languages. Trigonometry books in the United States in the early twentieth century were similar to this one. For most of the twentieth century, however, plane and spherical trigonometry were separated and only plane trigonometry was taught in schools.

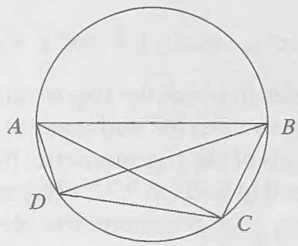
9.2.3 Problems

1. a. If A and B are points on a circle O of radius 1, prove that $\frac{1}{2}AB = \sin\frac{1}{2}(\angle AOB)$. (This equation shows how a table of chords in a circle can determine a table of sines.)

b. Does the formula of part a work if $\angle AOB$ is measured by its major rather than its minor arc? Why or why not?

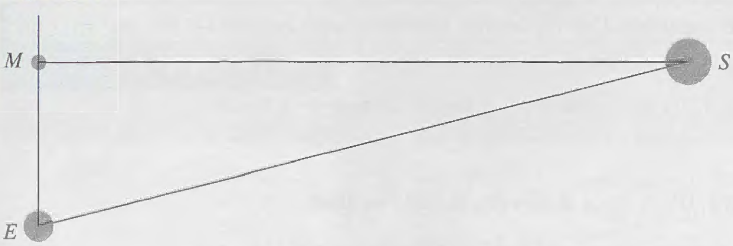
***2.** Ptolemy was the first to solve the following problem. Given two stars in the sky whose longitude and latitude on the celestial sphere are known, find the angle between them. Ptolemy sliced the celestial sphere by a plane through the two stars and the center of the earth, obtaining a great circle. If C and D represent the two stars, and the diameter of the circle is $\overline{AB}$, so that A, B, C, and D are in order on the circle (see Figure 29), then Ptolemy was able to prove, using similar triangles, that $AB \cdot CD + AD \cdot BC = BD \cdot AC$. Derive his theorem. (*Hint*: Choose point E on $\overline{AC}$ so that $\angle ABE \cong \angle EBC$. Then find two pairs of similar triangles.)

Figure 29



3. **Earth–Moon distance.** Aristarchus of Samos (c. 310–230 B.C.) used trigonometry to estimate the distance from Earth to the Sun. His procedure was to measure the angle $\angle MES$ between the Moon and the Sun at the moment when the moon was a half moon. (To exhibit the geometric setting, Figure 30 is not drawn to scale.) At that moment, $\angle SME$ was a right angle. It was difficult for Hipparchus (or for anyone else for that matter!) to determine when the Moon is exactly half full from observation on Earth. He estimated the angle $\angle MES$ to be about 87°. Because Hipparchus did not know the distance from Earth to the Moon, he expressed his estimate of the Earth–Sun distance as a multiple of the Earth–Moon distance.

Figure 30



a. Based on the estimate of 87° that Hipparchus made for $\angle MES$ and an Earth–Moon distance of 238,800 miles, find the corresponding estimate of the distance from Earth to the Sun in miles.

b. The actual average distance from Earth to the Sun is about 93 million miles. What is the corresponding degree measure of the angle $\angle MES$?

c. Explain why small errors in the measurement of $\angle MES$ result in relatively large errors in measurement of the Earth–Sun distance.

4. **Circumference of Earth.** Eratosthenes (275–194 B.C.) of Cyrene (a city in what is now Libya) was generally regarded as second only to Archimedes among mathematicians of his era. Among his many accomplishments, he was the first to measure the size of Earth on the basis of Earth observations. His procedure can be described as follows. At noon of the summer solstice (June 21), it was known that the Sun's rays struck the bottom of a deep well in the Egyptian town of Syene (now Aswan), due south of Alexandria. At the same time, Eratosthenes observed that the Sun made an angle of 7.2° to the vertical in Alexandria. Taking Earth to be a sphere, Eratosthenes concluded that the circumference of Earth was 50 times the distance between Alexandria and Syene.

a. Explain this conclusion.

b. Given that the surface distance from Alexandria to Syene is about 500 miles, what is the corresponding circumference and radius of Earth?

c. Suppose Eratosthenes was off by .2° in his observations. What then would be possibilities for his estimate for Earth's circumference?

Unit 9.3 Properties of the Sine and Cosine Functions

The trigonometric functions are among the most interesting functions in all of mathematics. An analysis of them using the ideas of Section 3.2.1 reveals a host of special properties. In this unit we separate those special properties into three types: algebraic, by which we mean the trigonometric identities that relate the values of these functions; geometric (or graphical), including examination of their graphs and their relationships to physical phenomena; and analytical, those properties that relate these functions and their derivatives and the consequent applications.

9.3.1 Algebraic properties of the trigonometric functions

Most courses in trigonometry place considerable emphasis on deriving and manipulating trigonometric identities. There are the *defining identities* that relate the basic values of the six trigonometric functions. For all values of t for which the denominators do not equal 0,

$$\tan t = \frac{\sin t}{\cos t}, \quad \cot t = \frac{\cos t}{\sin t}, \quad \sec t = \frac{1}{\cos t}, \quad \text{and} \quad \csc t = \frac{1}{\sin t}.$$

You have also seen the *Pythagorean identities*. For all t,

$$\sin^2 t + \cos^2 t = 1, \quad 1 + \tan^2 t = \sec^2 t, \quad \text{and} \quad 1 + \cot^2 t = \csc^2 t.$$

In Section 9.2.2, we discussed several examples in which the trigonometric functions are used to model periodic phenomena and to describe and analyze motion problems and geometric curves. Other applications of the trigonometric functions to the analysis of mechanical vibrations are discussed in Section 9.3.2. Trigonometric identities are at the heart of all such applications. The trigonometric identities govern and direct the use of the trigonometric functions in the same way that algebraic properties such as the commutative and distributive properties govern and direct the use of numbers.

The possible trigonometric identities are too numerous to list. In this section we content ourselves with a derivation of the most basic of the identities. Then in the Problems we ask you to use these to derive some of the other important identities.

A formula for rotation images

The definition of $\cos t$ and $\sin t$ as the x-coordinate and the y-coordinate of the image of $(1, 0)$ under a rotation centered at the origin of magnitude t makes it natural that trigonometric functions would be involved in formulas for rotation images in $\mathbf{R}^2$. Here is a proof of a theorem found in Section 7.2.2.

Theorem 9.9 **(Rotation Image Formula):** Let R_ϕ be the rotation with center $(0, 0)$ and magnitude ϕ. Then

$$R_\phi(x, y) = (x\cos\phi - y\sin\phi, x\sin\phi + y\cos\phi).$$

Proof: Because $R_\phi(1, 0) = (\cos\phi, \sin\phi)$, it follows that

$$R_\phi(x, 0) = (x\cos\phi, x\sin\phi), \text{ for all real numbers } x.$$

Also, $R_\phi(0, y) = (-y\sin\phi, y\cos\phi)$, for all real numbers y (see Problem 1). The vector $\mathbf{c} = (x, y)$ is the sum of the vectors $\mathbf{a} = (x, 0)$, and $\mathbf{b} = (0, y)$, and the points $(0, 0)$, $(x, 0)$, (x, y), and $(0, y)$ are successive vertices of a rectangle (see Figure 31). When the three vectors **a, b**, and **c** are rotated, the image of **c** is the sum of the images of **a** and **b**, since the image of a rectangle under a rotation is a congruent rectangle.

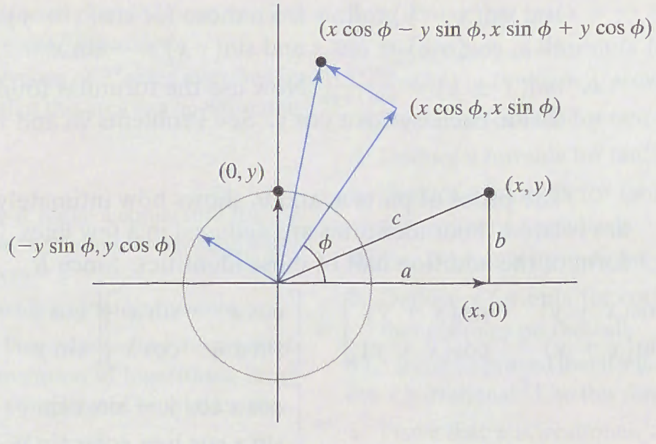
From the formulas, above,

$$R_\phi(\mathbf{a}) = R_\phi(x, 0) = (x\cos\phi, x\sin\phi)$$

and

$$R_\phi(\mathbf{b}) = R_\phi(0, y) = (-y\sin\phi, y\cos\phi).$$

Figure 31



So

$$R_\phi(\mathbf{c}) = R_\phi(x, y) = (x\cos\phi - y\sin\phi, x\sin\phi + y\cos\phi),$$

which completes the proof.

The product of the complex numbers $x + iy$ and $\cos\phi + i\sin\phi$ is $(x\cos\phi - y\sin\phi) + i(x\sin\phi + y\cos\phi)$. This once again shows that we can think of multiplying $x + iy$ by the complex number $\cos\phi + i\sin\phi$ as equivalent to applying R_ϕ to the point (x, y).

Formulas for $\cos(x \pm y)$ and $\sin(x \pm y)$

In Theorem 9.9, we used ϕ to identify the argument of the trigonometric functions because x and y are so identified with coordinates of points in $\mathbf{R}^2$. Now we use the Rotation Image Formula to derive identities involving the sum and difference of arguments. The letters ϕ and θ are sometimes used for these arguments, but more often these formulas are remembered using the letters x and y, as we show here. The symbol $\mp$, when used in conjunction with $\pm$, means that there are two sentences being written as one, with the top signs ($-$ and $+$) being used for one identity, and the bottom signs ($+$ and $-$) being used for the other.

Theorem 9.10 **(Sum and Difference Formulas):** For all real x and y, for which the expressions are defined

a. $\sin(x \pm y) = \sin x\cos y \pm \cos x\sin y$

b. $\cos(x \pm y) = \cos x\cos y \mp \sin x\sin y$

c. $\tan(x \pm y) = \dfrac{\tan x \pm \tan y}{1 \mp \tan x\tan y}$

Proof:

a. and **b.** (simultaneously!): Think of $(\cos(x + y), \sin(x + y))$ as the image of $(1, 0)$ under R_{x+y}. Then separate out the two rotations and apply the Rotation Image Theorem.

$$\begin{aligned}(\cos(x + y), \sin(x + y)) &= R_{x+y}(1, 0) && \text{(def. of cos and sin)}\\ &= R_x \circ R_y(1, 0) && \text{(angle addition)}\\ &= R_x(\cos y, \sin y)\\ &= (\cos x\cos y - \sin x\sin y, \sin x\cos y + \cos x\sin y)\end{aligned}$$

Equating the components of the first and last ordered pairs provides formulas for both $\cos(x + y)$ and $\sin(x + y)$. The corresponding formulas for $\cos(x - y)$

and $\sin(x-y)$ follow from those for $\cos(x+y)$ and $\sin(x+y)$ because for all x, $\cos(-x) = \cos x$ and $\sin(-x) = -\sin x$.

c. $\tan(x \pm y) = \frac{\sin(x\pm y)}{\cos(x\pm y)}$. Now use the formulas found in parts (a) and (b) and divide each by $\cos x \cos y$. See Problems 9a and 12.

The proof of parts **a.** and **b.** shows how intimately rotations, sines, and cosines are related. Four identities are deduced in a few lines. There is even a more elegant form of the addition half of these identities. Since $R_{x+y} = R_x \circ R_y$,

$$\begin{bmatrix} \cos(x+y) & -\sin(x+y) \\ \sin(x+y) & \cos(x+y) \end{bmatrix} = \begin{bmatrix} \cos x & -\sin x \\ \sin x & \cos x \end{bmatrix} \begin{bmatrix} \cos y & -\sin y \\ \sin y & \cos y \end{bmatrix}$$

$$= \begin{bmatrix} \cos x \cos y - \sin x \sin y & -\sin x \cos y + \cos x \sin y \\ \sin x \cos y - \cos x \sin y & \cos x \cos y - \sin x \sin y \end{bmatrix}.$$

The sum and difference formulas are significant because from them so many other formulas can be derived. By letting $x = y$, the *double-angle formulas* for $\cos 2x$ and $\sin 2x$ follow. From the double-angle formulas, formulas for $\cos 3x$, $\sin 3x$, $\cos 4x$, $\sin 4x$, etc., can be derived. Also, from the double-angle formulas, formulas for $\cos\left(\frac{x}{2}\right)$ and $\sin\left(\frac{x}{2}\right)$ can be deduced. By adding or subtracting the formulas for $\cos(x+y)$ and $\cos(x-y)$ and other pairs of sums and/or differences, the *product-to-sum identities* can be obtained. Before logarithms were discovered, the product-to-sum identities were used by mathematicians to perform difficult multiplications. These ideas are explored in the Problems.

9.3.1 Problems

1. Explain why, for all real numbers y, the image of $(0, y)$ under a rotation of magnitude ϕ around the origin is $R_\phi(0, y) = (-y\sin\phi, y\cos\phi)$.

2. Use the Pythagorean identities and the Sum Formulas to derive the following **double-angle formulas**.

a. $\cos(2x) = \cos^2 x - \sin^2 x$

b. $\cos(2x) = 1 - 2\sin^2 x$

c. $\cos(2x) = 2\cos^2 x - 1$

d. $\sin(2x) = 2\sin x \cos x$

3. Use the identities in Problem 2 to derive the following **half-angle formulas**.

a. Prove: For all x, $\left|\cos\left(\frac{x}{2}\right)\right| = \sqrt{\frac{1+\cos x}{2}}$.

b. Find and prove a formula for $\left|\sin\left(\frac{x}{2}\right)\right|$ in terms of $\cos x$.

c. Your formulas and proofs for parts **a** and **b** should be similar. Explain why there is no similar formula and proof for a formula for $\left|\sin\left(\frac{x}{2}\right)\right|$ in terms of $\sin x$.

4. Use the Sum Formulas to find **multiple-angle formulas**.

a. Find a formula for $\cos(3x)$ in terms of $\cos x$ and $\sin x$, and verify your formula with a specific value of x.

b. Find a formula for $\cos(3x)$ in terms of $\cos x$.

c. Find a formula for $\sin(3x)$ in terms of $\sin x$ and $\cos x$, and verify your formula with a specific value of x.

d. Find a formula for $\cos(4x)$ in terms of $\cos x$ and $\sin x$.

e. Find a formula for $\cos(4x)$ in terms of $\cos x$.

f. Find a formula for $\sin(4x)$ in terms of $\sin x$ and $\cos x$.

g. Prove that $\cos(nx)$ can be expressed as a polynomial of degree n in $\cos x$.

5. a. Derive an expression for $\frac{\sin 6x}{\sin 2x}$ in terms of $\cos x$.

b. Generalize part **a** in some way.

6. Use the fact that $\sin\left(\frac{\pi}{4}\right) = \frac{\sqrt{2}}{2} = \cos\left(\frac{\pi}{4}\right)$ and identities in Problem 3 to determine the following sines and cosines.

a. $\sin\left(\frac{\pi}{8}\right) = \frac{\sqrt{2-\sqrt{2}}}{2}$; $\cos\left(\frac{\pi}{8}\right) = \frac{\sqrt{2+\sqrt{2}}}{2}$

b. $\sin\left(\frac{\pi}{16}\right) = \frac{\sqrt{2-\sqrt{2+\sqrt{2}}}}{2}$; $\cos\left(\frac{\pi}{16}\right) = \frac{\sqrt{2+\sqrt{2+\sqrt{2}}}}{2}$

c. $\sin\left(\frac{\pi}{32}\right) = \frac{\sqrt{2-\sqrt{2+\sqrt{2+\sqrt{2}}}}}{2}$; $\cos\left(\frac{\pi}{32}\right) = \frac{\sqrt{2+\sqrt{2+\sqrt{2+\sqrt{2}}}}}{2}$

d. Do you think that the pattern established in parts **a**–**c** persists? Why or why not?

7. In 1593, Vieté proved that

$$\frac{2}{\pi} = \sqrt{\frac{1}{2}} \cdot \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2}}} \cdot \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2}}}} \cdot \ldots.$$

He found this formula by computing areas of regular polygons with $4, 8, 16, \ldots, 2^n$ sides inscribed in a circle of radius 1. This problem asks you to carry out the steps of this derivation.

a. Find the area of a regular polygon of 2^2 sides inscribed in a circle of radius 1.

b. Find the area of a regular polygon of 2^3 sides inscribed in a circle of radius 1, and show that this area can be written in the form $\frac{2}{\sqrt{\frac{1}{2}}}$.

c. Show that as the number of sides is doubled from n to $2n$, an area equal to $n\left(\sin\frac{\pi}{n}\right)\left(1-\cos\frac{\pi}{n}\right)$ is added.

d. Find the area of a regular polygons of 2^4 sides inscribed in a circle of radius 1, and show that this area can be written in the form $\frac{2}{\sqrt{\frac{1}{2}}\cdot\sqrt{\frac{1}{2}+\frac{1}{2}\sqrt{\frac{1}{2}}}}$.

e. Based on a pattern in parts **a–d**, make a conjecture for the area of a regular polygon of $n = 2^k$ sides inscribed in a circle of radius 1.

f. Show how your conjecture leads to Vieté's formula.

8. Use the Sum and Difference Formulas to find the **product-to-sum formulas**. Before the invention of logarithms, these formulas were used to perform complicated multiplications.

a. Prove that $\cos x \cos y = \frac{1}{2}(\cos(x+y) + \cos(x-y))$.

b. Assume you have a table of cosines (use a calculator for this) that enables you to find inverse cosines as well. Multiply 95632 by 61807 in the following way. Determine $x = \cos^{-1}(.95632)$ and $y = \cos^{-1}(.61807)$ with your "table". Find $\cos(x+y)$ and find $\cos(x-y)$, again using your table. Divide by 2 and put the decimal point in the proper place.

c. Prove that $\sin x \sin y = \frac{1}{2}(\cos(x-y) - \cos(x+y))$ and perform the same multiplication as in part **b** using that identity.

9. a. Prove that $\tan(x \pm y) = \frac{\tan x \pm \tan y}{1 \mp \tan x \tan y}$.

b. By dividing the formula for $\sin(2x)$ by the formula for $\cos(2x)$ in Problem 2, prove that $\tan(2x) = \frac{2\tan x}{1-\tan^2 x}$.

c. Deduce a formula for $\tan(3x)$ in terms of $\tan x$.

d. Deduce a formula for $\tan\left(\frac{x}{2}\right)$ in terms of $\cos x$.

e. Deduce a formula for $\tan\left(\frac{x}{2}\right)$ in terms of $\sin x$ and $\cos x$ that contains no radicals.

10. a. Deduce a formula for $\cot(2x)$ in terms of $\cot x$.

b. Deduce a formula for $\cot\left(\frac{x}{2}\right)$ in terms of $\sin x$ and $\cos x$ that contains no radicals.

11. It can be proved that if x is a nonzero rational number, then $\cos x$ is irrational.[3] Use this result in the following problems.

a. Prove that π is irrational.

b. Prove that $\sin x$, $\tan x$, $\cot x$, $\sec x$, and $\csc x$ are irrational whenever x is a nonzero rational number for which the function is defined. (*Hint*: Use formulas for $\cos 2x$.)

c. Prove that nonzero values of $\sin^{-1} x$, $\cos^{-1} x$, and $\tan^{-1} x$ are irrational whenever x is a rational number in their domains.

12. Use a formula for $\tan(x-y)$ (see Problem 9**a**) to deduce a formula for the tangent of the acute angle formed by two non-vertical and nonperpendicular lines with slopes m_1 and m_2.

9.3.2 Geometric properties of the sine and cosine functions

We know that, by definition, the sine and cosine functions are *periodic with period* 2π. That is, if $p = 2\pi$, then

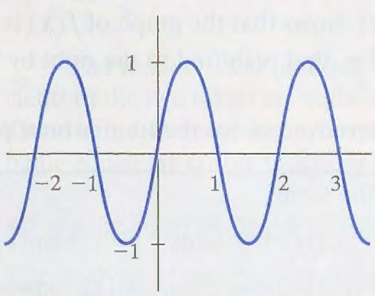
$$\sin(x+p) = \sin(x) \quad \text{and} \quad \cos(x+p) = \cos(x), \qquad \text{for all real numbers } x,$$

and no smaller positive value of p has this property. More generally, for each positive number c, the functions

$$x \to \sin(cx) \quad \text{and} \quad x \to \cos(cx)$$

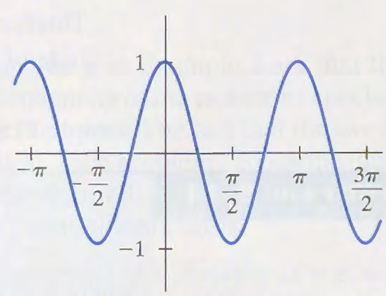
have period $p = \frac{2\pi}{c}$. In some applications, the period p is called the **wave length**. We also say that these functions have **frequency** $f = \frac{c}{2\pi}$. Thus, for example, the real function $x \to \sin(\pi x)$ has period 2 and frequency $\frac{1}{2}$ (see Figure 32), while $x \to \cos(2x)$ has period π and frequency $\frac{1}{\pi}$ (see Figure 33).

Figure 32



$y = \sin(\pi x)$

Figure 33



$y = \cos(2x)$

[3]For a proof, see Ivan Niven, *Irrational Number*, Carus Monograph No. 11 (Washington, DC: Mathematical Association of America and John Wiley, 1956), 17.

Sum of multiples of a sine and a cosine function with the same period

Combinations of sine and cosine functions often have interesting but predictable graphical features, as the examples in this section illustrate. It is best if you have a graphing utility as you read this section, so that you can construct the graphs shown here while you read.

EXAMPLE 1 **a.** Graph the function f defined by

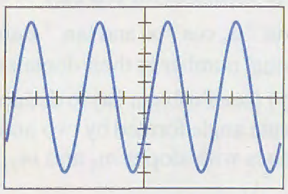
$$f(x) = -3\cos(2x) + 4\sin(2x)$$

on the window $-2\pi < x < 2\pi$, $-6 < y < 6$.

b. Analyze why the graph looks the way that it does.

Solution **a.** Your graph should look like the graph of a positive multiple of the sine or cosine function that has been shifted along the x-axis (see Figure 34). The values of the function appear to range from -5 to 5 and the period should appear to be about π, which is the same as the period of the given sine and cosine terms.

Figure 34



$-2\pi < x < 2\pi$, x-scale $= 1$
$-6 < y < 6$, y-scale $= 1$

b. It seems that $f(x)$ is expressible in the form $f(x) = A\cos(c[x - B])$ or $f(x) = A\sin(c[x + B])$ for appropriate constants A and B. Either of these forms of $f(x)$ would explain the apparent nature of the graph of $f(x)$ shown in Figure 34. To obtain the $A\cos(c[x - B])$ form, we multiply and divide the given expression for $f(x)$ by $5 = \sqrt{(-3)^2 + 4^2}$ to obtain

$$f(x) = -3\cos(2x) + 4\sin(2x) = 5\left(\frac{-3}{5}\cos(2x) + \frac{4}{5}\sin(2x)\right).$$

We have chosen the multiplier so that the coefficients, $-\frac{3}{5}$ and $\frac{4}{5}$ are coordinates of a point $\left(-\frac{3}{5}, \frac{4}{5}\right)$ on the unit circle. Consequently, there is an angle θ such that

$$\cos\theta = -\frac{3}{5} \quad \text{and} \quad \sin\theta = \frac{4}{5}.$$

(This angle has measure approximately 2.214 radians or 126.9°.) Therefore, we can express $f(x)$ in the form

$$\begin{aligned} f(x) &= 5(\cos\theta\cos(2x) + \sin\theta\sin(2x)) \\ &= 5\cos(2x - \theta) \\ &= 5\cos\left[2\left(x - \frac{\theta}{2}\right)\right]. \end{aligned}$$

This form of $f(x)$ shows that the graph of $f(x)$ is a cosine curve with an amplitude of 5, period π, that is shifted to the right by $\frac{\theta}{2}$.

Example 1 can be generalized into a theorem, whose proof we leave to you.

Theorem 9.11 Every function f of the form

$$f(x) = a\cos(cx) + b\sin(cx),$$

where a, b, and c are real numbers, can also be expressed in the form

$$f(x) = A\cos[c(x - B)],$$

where $A = \sqrt{a^2 + b^2}$, $\cos(Bc) = \frac{a}{A}$, and $\sin(Bc) = \frac{b}{A}$.

Either of the forms $f(x) = A\cos[c(x - B)]$ or $f(x) = A\sin[c(x + B)]$ is called a **phase-amplitude form** of $f(x)$ because the constant A represents the *amplitude* and B the horizontal shift or *phase shift* from the standard position of the corresponding sine or cosine function. As you will see later, the phase-amplitude form of a function is quite useful in analyzing vibrational phenomena.

Sums of multiples of two sine and/or cosine functions with the same amplitude and different periods

When multiples of sine and cosine functions have different periods, adding them may result in a function whose graph is not a pure sine wave. Still, a simple equation for the sum can be obtained and its graph analyzed.

EXAMPLE 2 **a.** Graph the function g defined by $g(x) = 2\cos(8x) - 2\cos(10x)$ on the window

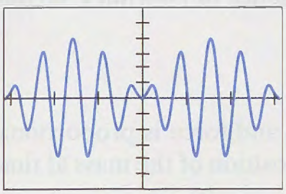
$$-\pi < x < \pi, -6 < y < 6.$$

b. Analyze why the graph looks the way that it does.

Solution **a.** Your graph should look like that of a sine or cosine function with a small period that has been "pinched" along its length near $x = -\pi$, $x = 0$ and $x = \pi$ (see Figure 35). The graph appears to be symmetric to the y-axis.

If you graph this function over a larger interval, for x such as $-4\pi < x < 4\pi$, you may find that the graph breaks up badly due to the pixel limitations of your calculator screen, but you should still see that the pinching effect persists in intervals of apparent length 2π.

Figure 35



$-\pi < x < \pi$, x-scale = 1
$-6 < y < 6$, y-scale = 1

b. Note that the period of the first term is $\frac{2\pi}{8} = \frac{\pi}{4}$ while that of the second term is $\frac{2\pi}{10} = \frac{\pi}{5}$. Also note that

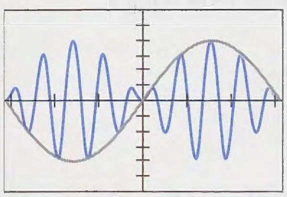
$$g(x) = 2\cos(8x) - 2\cos(10x)$$
$$= 2\cos(9x - x) - 2\cos(9x + x).$$

This is in the form of the identity of Problem 8**c** in Section 9.3.1. From that identity,

$$g(x) = 4\sin(9x)\sin x.$$

This expression for $g(x)$ reveals why its graph has its unusual appearance. The $4\sin(9x)$ term is periodic with period $\frac{2\pi}{9}$ and amplitude 4. It is multiplied by the factor $\sin x$, which has period 2π, and has values ranging from -1 to 1. Thus, the graph of the $h(x) = \sin x$ factor provides an outline for the graph of g. This is seen in Figure 36.

Figure 36



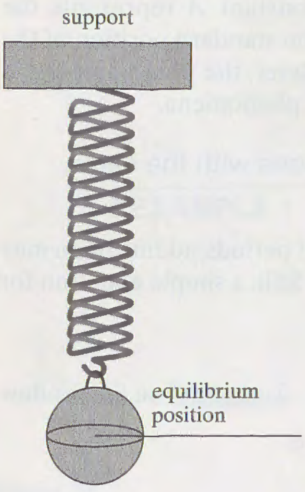
$-\pi < x < \pi$, x-scale = 1
$-6 < y < 6$, y-scale = 1

The most critical features of the function g in Example 2 are that the coefficients of the two terms are equal and the frequencies of the two terms are close in size. Thus their difference is much smaller than their sum. The fact that the two functions being added are cosine functions is not critical. (See Problem 7 following this section.)

An application to the analysis of mechanical vibrations.

The analysis of mechanical vibrations is important to a diversity of real-world phenomena, from situations as large as the design of machinery and earthquake-resistant structures to the analysis of atomic and molecular vibrations, among others. As with many applications of mathematics, there is a simple physical model that is used as a basis for mathematical models of the application. For example, in probabilistic or

Figure 37



statistical applications, the physical model of drawing cards from a deck or balls from an urn is a simple but useful model. For problems in mechanical vibrations, the *spring-mass* system pictured in Figure 37 is useful.

When the mass at the base of the spring (shown in Figure 37 as a sphere) is disturbed in the vertical direction at a certain time $t = 0$, by positioning it above or below the equilibrium position and/or by imparting an initial velocity upward or downward, the mass will oscillate in the vertical direction for $t > 0$. If $x(t)$ is the directed distance of the center of the mass from its equilibrium position at time $t \geq 0$, then $x(t)$ depends not only on the initial position $x_0 = x(0)$ and the initial velocity $v_0 = x'(0)$ given to the mass but also on the mass m and the physical characteristics of the spring.

Experience suggests that the spring oscillates and, in the idealized absence of friction, would oscillate forever. But it is a wonderful surprise that the height of the mass is given by a sine or cosine function. Here is why.

Within their elastic limits, stretching and compressing physical springs typically satisfy Hooke's Law: *The magnitude $|F|$ of the force F that is required to stretch or compress a spring a distance D is proportional to D; that is, $|F| = kD$ for some positive constant k.* The constant k is called the **spring constant** for the spring. Springs with large spring constants are stiff; those with small spring constants stretch and compress more easily.

Suppose that a spring-mass system consists of a mass m suspended from a spring with spring constant k. If the internal friction of the spring and other forces acting on the spring are assumed to be negligible, then as a consequence of **Newton's Second Law of Motion** $ma = F$; that is,

$$(\text{mass})(\text{acceleration}) = \text{force}.$$

Because acceleration is the second derivative of position, and force is proportional to the second derivative, the function $x(t)$ describing the position of the mass at time t must satisfy the differential equation

$$(1) \qquad m\frac{d^2x(t)}{dt^2} = -kx(t) \qquad \text{for all } t \geq 0$$

and the initial conditions:

$$(2) \qquad x(0) = x_0 \qquad \text{(initial position)}$$

$$(3) \qquad \frac{dx(0)}{dt} = v_0 \qquad \text{(initial velocity).}$$

Equations (1), (2), and (3) are sometimes called the **equations of motion** for the spring-mass system.

EXAMPLE 3 Suppose that suspending a 6-lb weight from a certain spring stretches it 6 in. At time $t = 0$, the weight is moved 4 in. above its equilibrium position and given an initial velocity of 2 ft per second upward. Find the equations of motion of this spring-mass system.

Solution In the ft-lb-sec system, mass is measured in slugs, and mass m in slugs of an object is related to its weight w in pounds by $w = 32m$. Also, because the spring is stretched 6 in. $= \frac{1}{2}$ ft, the spring constant k is given by 6 lb $= k \cdot \frac{1}{2}$ ft or $k = 12$ lb per ft. Therefore, the directed distance $x(t)$ of the weight above the equilibrium position satisfies the equation $\frac{6}{32}\frac{d^2x(t)}{dt^2} = -12x(t)$, which simplifies to

$$(1') \qquad \frac{d^2x(t)}{dt^2} = -64x(t) \qquad \text{for all } t \geq 0.$$

By adding the information about the initial position and velocity, we obtain (2′) and (3) of the equations of motion.

(2′) $$x(0) = \frac{1}{3}\text{ ft} \qquad \text{(initial position)}$$

(3′) $$\frac{dx(0)}{dt} = 2\text{ ft per sec} \qquad \text{(initial velocity)}$$

It is possible to find the motion function $x(t)$ from the equations of motion of the spring-mass system in Example 3. This requires, quite obviously, use of the formulas for differentiation of the sine and cosine functions derived in calculus. Recall that if $f(t) = \sin t$, then $f'(t) = \frac{df(t)}{dt} = \cos t$, and if $f(t) = \cos t$, $f'(t) = \frac{df(t)}{dt} = -\sin t$. From this, we can check that some specific functions satisfy the equations of motion.

Question 1: Verify that each of the functions

$$x_1(t) = \cos\left[\sqrt{\frac{k}{m}}t\right] \quad \text{and} \quad x_2(t) = \sin\left[\sqrt{\frac{k}{m}}t\right]$$

satisfy equation (1) of the equations of motion. More generally, show that if a and b are any real numbers, the function

$$x_3(t) = a\cos\left(\sqrt{\frac{k}{m}}t\right) + b\sin\left(\sqrt{\frac{k}{m}}t\right) \qquad \text{for all } t \geq 0$$

satisfies equation (1).

Question 2: Show that if $a = x_0$ and $b = v_0\sqrt{\frac{m}{k}}$ in the function x_3 of Question 1, which satisfies equation (1) of the equations of motion, the resulting function also satisfies equations (2) and (3).

EXAMPLE 4 Consider the spring-mass system described in Example 3.

a. Use the results of Question 2 to find the motion function $x(t)$ from the equations of motion of the system.

b. Write $x(t)$ in phase-amplitude form and use it to describe the graphical characteristics of the motion function $x(t)$ of the system.

c. Use a graphing utility to confirm the conclusions from part (b).

Solution **a.** $x(t) = \cos x_0\left(\sqrt{\frac{m}{k}}t\right) + v_0\sqrt{\frac{m}{k}}\sin\left(\sqrt{\frac{m}{k}}t\right)$. Here $k = -12\frac{lb}{ft}$ and $m = \frac{6}{32}$ slugs. Thus

$$x(t) = \frac{1}{3}\cos(8t) + \frac{1}{4}\sin(8t).$$

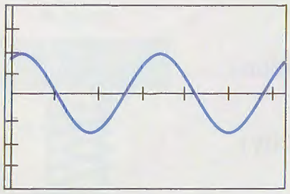
b. To write $x(t)$ in the phase-amplitude form $A\cos[8(t - B)]$, note that

$$A = \sqrt{\left(\frac{1}{3}\right)^2 + \left(\frac{1}{4}\right)^2} = \frac{5}{12}; \quad \cos(8B) = \frac{\frac{1}{3}}{\frac{5}{12}} = \frac{4}{5}; \quad \sin(8B) = \frac{\frac{1}{4}}{\frac{5}{12}} = \frac{3}{5}$$

and so $8B \approx .6435$ radians or 36.8°. Therefore, $B \approx .08$ radians or 4.61°. Therefore, a phase-amplitude form of $x(t)$ is

$$x(t) \approx \frac{5}{12}\cos[8(t - .08)],$$

Figure 38



$0 < t < \frac{\pi}{2}$, t-scale = 0.25
$-1 < x < 1$, x-scale = 0.25

which shows that the amplitude of the oscillation of the mass is $\frac{5}{12}$ ft or 5 in. from the equilibrium position.

c. In the window $0 < t < \frac{\pi}{2}, -1 < x < 1$, the graph of $x(t)$ looks like the graph in Figure 38. This graph shows that the mass, which is initially 4 in. above its equilibrium, continues upward for another inch because the initial velocity is upward, before beginning its downward movement toward its equilibrium. The frequency of the motion is $\frac{8}{2\pi}$ seconds or 1.27 seconds, so the motion pictured in Figure 38 takes place in a little more than 2.5 seconds.

In the preceding discussion, it was assumed that no outside forces and no damping forces were acting on the spring-mass system. If an outside force is acting on the system, then the resulting motion would obviously be affected. For example, suppose that an outside force is acting on the spring-mass system in Example 1 so that it is itself vibrating according to the function

$$f(t) = \frac{1}{4}\sin(2\pi t).$$

(This might be accomplished by moving the support up and down with an amplitude of $\frac{1}{4}$ ft starting at $t = 0$.) Then it can be shown that the motion function $x(t)$ of the mass would be a function of the form

$$x(t) = A\cos[8(t - B)] + C\sin[2\pi t],$$

where A, B, and C are constants determined by the initial conditions, the mass m and the spring constant k. This is the sort of function that we considered in Examples 3 and 4. Consequently, a motion of the sort displayed in Example 3 may result. In the analysis of mechanical vibrations, such motions are called **beats**. Most of us have experienced beats in sound vibrations. For example, the sound that we hear from an electric motor running at a constant speed might have a periodically varying amplitude because the support on which the motor is mounted is vibrating at frequency close to the frequency of the sound produced by the motor itself.

9.3.2 Problems

1. Show that the function f with $f(t) = -3\cos(2t) + 4\sin(2t)$ in Example 1 can also be expressed as $f(t) = A\sin[c(t + B)]$.

2. Prove Theorem 9.11.

3. Prove the analogue to Theorem 9.11 suggested by Problem 1 of this set.

4. Graph the function g with $g(x) = \sin x + \cos x$, and explain why the graph has the shape it has.

5. Graph $h(x) = 7\cos x - 24\sin x$ and explain why the graph has the period and amplitude it has.

6. a. Graph $f(x) = \sin x - \cos x$ and explain why its graph looks the way it does.

b. Repeat part **a** for the function $g(x) = (\sin x - \cos x)e^{-x}$.

7. Suppose that c_1 and c_2 are real numbers such that $c_1 > c_2 > 0$ such that $c_1 + c_2$ is much larger than $c_1 - c_2$. Prove: If a is any nonzero real number and if $f(x)$ has any of the following forms,

$$f(x) = \begin{cases} a[\cos(c_1x) \pm \cos(c_2x)] \\ a[\sin(c_1x) \pm \cos(c_2x)] \\ a[\cos(c_1x) \pm \sin(c_2x)] \\ a[\sin(c_1x) \pm \sin(c_2x)] \end{cases}$$

then f has a graph somewhat like that of the function in Example 2 in this section.

ANSWERS TO QUESTIONS

1. If $x(t) = \cos\left(\sqrt{\frac{k}{m}}t\right)$, then $x'(t) = -\sqrt{\frac{k}{m}}\sin\left(\sqrt{\frac{k}{m}}t\right)$, and $x''(t) = -\frac{k}{m}\cos\left(\sqrt{\frac{k}{m}}t\right)$, shows that equation (1) holds. If $x(t) = \sin\left(\sqrt{\frac{k}{m}}t\right)$, then $x'(t) = \sqrt{\frac{k}{m}}\cos\left(\sqrt{\frac{k}{m}}t\right)$, and $x''(t) = -\frac{k}{m}\sin\left(\sqrt{\frac{k}{m}}t\right)$, so equation (1) again follows. If $x(t) = a\cos\left(\sqrt{\frac{k}{m}}t\right) + b\sin\left(\sqrt{\frac{k}{m}}t\right)$, then $x'(t) = -a\sqrt{\frac{k}{m}}\sin\left(\sqrt{\frac{k}{m}}t\right) + b\sqrt{\frac{k}{m}}\cos\left(\sqrt{\frac{k}{m}}t\right)$, so $x''(t) = -a\frac{k}{m}\cos\left(\sqrt{\frac{k}{m}}t\right) - b\frac{k}{m}\sin\left(\sqrt{\frac{k}{m}}t\right)$, and again equation (1) holds.

2. $x_3(t) = x_0\cos\left(\sqrt{\frac{k}{m}}t\right) + v_0\sqrt{\frac{m}{k}}\sin\left(\sqrt{\frac{k}{m}}t\right)$, so $x_3(0) = x_0\cos\left(\sqrt{\frac{k}{m}}\cdot 0\right) + v_0\sqrt{\frac{m}{k}}\sin\left(\sqrt{\frac{k}{m}}\cdot 0\right) = x_0\cdot 1 + 0 = x_0$, so equation (2) is satisfied. $\frac{dx(0)}{dt} = \frac{dx_3(0)}{dt} = -x_0\sqrt{\frac{k}{m}}\sin\left(\sqrt{\frac{k}{m}}\cdot 0\right) + v_0\sqrt{\frac{m}{k}}\cdot\sqrt{\frac{k}{m}}\cos\left(\sqrt{\frac{k}{m}}\cdot 0\right) = v_0$. So equation (3) is satisfied.

9.3.3 Analytical properties of the sine and cosine functions

The formulas that you learned in calculus for the derivatives of the sine and cosine functions, for x in radians,

$$\frac{d}{dx}(\sin x) = \cos x \quad \text{and} \quad \frac{d}{dx}(\cos x) = -\sin x,$$

and the corresponding second derivative formulas

$$\frac{d^2}{dx^2}(\sin x) = -\sin x \quad \text{and} \quad \frac{d^2}{dx^2}(\cos x) = -\cos x,$$

are remarkable in their simplicity and yet powerful in a variety of applications in pure and applied mathematics.

In calculus, these formulas are often derived analytically by applying the limit definition of the derivative, the addition formulas for the sine and cosine, and the special limit:

$$\lim_{\theta\to 0}\frac{\sin\theta}{\theta} = 1 \qquad \theta \text{ in radians} \qquad \text{(see Problem 3).}$$

Although these derivations are straightforward, they give little insight into the nature of these functions as tools for modeling periodic behavior.

We now show that analytic properties of the sine and cosine functions embodied in the differentiation formulas can be understood dynamically by using the sine and cosine functions as models of periodic behavior. Specifically, we will explain these formulas in terms of a mathematical model for *uniform circular motion in a plane.*

Uniform circular motion of an object in a plane

A uniform circular motion model is appropriate for physical situations such as (a) the motion of a communications satellite in a circular orbit around earth, or (b) motion of an object that you are spinning rapidly in a circle by holding one end of a string with the other end tied to the object. A precise description of uniform circular motion of an object requires statements about its path, its velocity, and its speed.

(P) Path of the object: The object travels in a path T that is a directed circle (i.e., a circle with a specified direction) with a radius r and with a center at a point O in that plane.

(V) Velocity and speed of the object: The speed s of the object along the path is constant, and the velocity vector of the object at any point P of the path is tangent to the path T at P and points in the direction of the path.

We need to say more than this to describe the situations (a) and (b) mentioned above. For instance, if the string in situation (b) would break, the object would fly off in a direction tangent to the circle at the point P where the object is located when the string breaks. Your hand must exert (centripetal) force to keep the object moving along the circular path T. This force F is exerted on the object in the direction from the object to your hand and follows Newton's law $F = ma$. With this example in mind, we can add the following statement to the description of uniform motion.

(A) Acceleration of the object: The object is accelerated toward the center of the circular path T and the magnitude of the acceleration is constant. (The centripetal force on the object accounts for the acceleration.)

We assume that all forces other than the centripetal force are negligible relative to the centripetal force.

Mathematical models of uniform circular motion

Now we create a mathematical model for uniform circular motion. First, we need a convenient location and description for the directed path T. Since T is to travel along a circle of radius R, suppose that we locate that circle in the xy-plane with its center at the origin. Then the path T is located on the circle

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

However, this equation is not a suitable description of the path. Motion means that the position P of the object depends on time. The equation $x^2 + y^2 = R^2$ does not allow us to determine the position of the object at a given time.

What we need is a description of the path in which the time t is a parameter and that gives the location of the point P along the circular path at any time t. This can be done with an equation for the position vector $\mathbf{r(t)}$ joining O to P at time t,

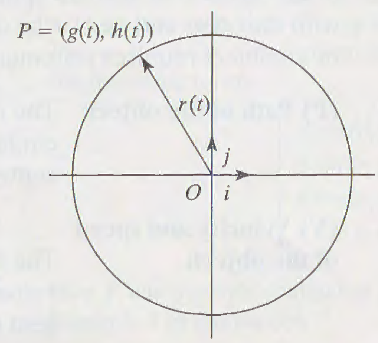
$$\mathbf{r(t)} = g(t)\mathbf{i} + h(t)\mathbf{j},$$

where $\mathbf{i} = (1, 0)$ and $\mathbf{j} = (0, 1)$ are the standard unit vectors, and $g(t)$ and $h(t)$ are real functions defined for all real numbers t. Any such path has a *natural direction*—the direction of increasing t values—and a *natural initial point*—the point that corresponds to $t = 0$.

We now identify conditions on the functions $g(t)$ and $h(t)$ that correspond to the path, velocity, and acceleration requirements (P), (V), and (A) above for uniform circular motion on a circle of radius R centered at the origin O. Recall that velocity is defined as the derivative of position with respect to time, and the derivative of velocity with respect to time is acceleration. Thus, since the position vector $\mathbf{r}$ joining O to P (Figure 39) is given by the vector function

$$\mathbf{r(t)} = g(t)\mathbf{i} + h(t)\mathbf{j},$$

Figure 39



the velocity vector **v** at P is

$$\mathbf{v(t)} = g'(t)\mathbf{i} + h'(t)\mathbf{j}.$$

The speed $s(t)$ at time t is the magnitude of the velocity vector. That is,

$$s(t) = |\mathbf{v(t)}| = \sqrt{g'(t)^2 + h'(t)^2}.$$

The acceleration vector **a** is given by

$$\mathbf{a(t)} = g''(t)\mathbf{i} + h''(t)\mathbf{j}.$$

To assure that this path is on the circle of radius R centered at the origin as required in condition (P), we need to require that $g(t)$ and $h(t)$ satisfy the two conditions shown here as (P*).

(P*) **i.** $g(t)^2 + h(t)^2 = R^2$ for all t, and
ii. for all (x, y) with $x^2 + y^2 = R^2$, there is a t such that $x = g(t)$ and $y = h(t)$.

According to the velocity condition (V), the speed of the object must be constant for uniform circular motion. Consequently, $g(t)$ and $h(t)$ must also satisfy the velocity condition (V*).

(V*) The functions $g(t)$ and $h(t)$ are differentiable and there is a constant K such that $g'(t)^2 + h'(t)^2 = K^2$ for all t.

Condition (V) for uniform circular motion also requires that the velocity vector **v(t)** at any point P along the circular path T must be perpendicular to the position vector **v(t)** at P. (The tangent to a circle is perpendicular to the radius at its endpoint on the circle.) In Problem 4, you are asked to verify that $\mathbf{r(t)} = g(t)\mathbf{i} + h(t)\mathbf{j}$ and $\mathbf{v(t)} = g'(t)\mathbf{i} + h'(t)\mathbf{j}$ are perpendicular vectors if and only if the following condition is satisfied,

$$g(t)g'(t) + h(t)h'(t) = 0 \text{ for all } t;$$

that is, if their dot product is zero. However, this condition is automatically satisfied for differentiable functions $g(t)$ and $h(t)$ that satisfy condition (P) because

$$g(t)^2 + h(t)^2 = R^2 \text{ for all } t \implies 0 = \frac{d}{dt}(R^2) = 2g(t)g'(t) + 2h(t)h'(t) \text{ for all } t.$$

Consequently, conditions (P*) and (V*) together assure that the requirements (P) and (V) for uniform circular motion on a circle of radius R are satisfied.

Finally, we consider condition (A) on the motion. This requires that the acceleration of the object is directed toward the center of the circular path and that it has a constant magnitude. Because the acceleration is directed toward the center of the circle, its direction is opposite to that of the position vector $\mathbf{r(t)} = g(t)\mathbf{i} + h(t)\mathbf{j}$. Because its magnitude is constant, it is simply a negative multiple of the position vector. Consequently, requirement (A) for uniform circular motion is met if the functions $g(t)$ and $h(t)$ satisfy the following condition.

(A*) There is a positive constant M such that $\mathbf{a(t)} = -M\mathbf{r(t)}$; that is,

$$g''(t) = -Mg(t) \quad \text{and} \quad h''(t) = -Mh(t) \text{ for all } t.$$

In summary, if we can find a pair of twice differentiable functions $g(t)$ and $h(t)$ that satisfy conditions (P*), (V*), and (A*), then the vector function

$$\mathbf{r(t)} = g(t)\mathbf{i} + h(t)\mathbf{j}$$

provides a mathematical model for uniform circular motion.

Notice that the requirements

$$g(t)^2 + h(t)^2 = R^2 \quad \text{and} \quad g'(t)^2 + h'(t)^2 = K^2 \qquad \text{for all } t$$

in (P*) and (V*) bear a strong resemblance to the Pythagorean identity $\cos^2 x + \sin^2 x = 1$, and that the requirement

$$g''(t) = -Mg(t) \quad \text{and} \quad h''(t) = -Mh(t) \qquad \text{for all } t$$

is reminiscent of the derivative formulas

$$\frac{d^2}{dx^2}(\sin x) = -\sin x \quad \frac{d^2}{dx^2}(\cos x) = -\cos x \qquad x \text{ in radians.}$$

These similarities suggest that vector function

$$\mathbf{r(t)} = \mathrm{R}(\cos t)\mathbf{i} + \mathrm{R}(\sin t)\mathbf{j}$$

might satisfy all the conditions (P*), (V*), and (A*). This is the case.

Theorem 9.12(a) The vector function

$$\mathbf{r(t)} = R\cos(t)\mathbf{i} + R\sin(t)\mathbf{j}$$

describes uniform circular motion for an object P on a circle of radius R centered at the origin O with a constant speed R and a constant magnitude of acceleration R.

Proof: Condition (P*) can be verified for this position function as follows:

i. $\sqrt{(R\cos t)^2 + (R\sin t)^2} = \sqrt{R^2(\cos^2 t + \sin^2 t)} = R$,

ii. If (x, y) satisfies $x^2 + y^2 = R^2$, then $\left(\frac{x}{R}, \frac{y}{R}\right)$ is a point on the unit circle and so $\frac{x}{R} = \cos t, \frac{y}{R} = \sin t$ for some t with $0 \le t < 2\pi$. Since

$$\frac{d}{dx}(\sin x) = \cos x \quad \text{and} \quad \frac{d}{dx}(\cos x) = -\sin x,$$

and

$$\frac{d^2}{dx^2}(\sin x) = -\sin x \quad \text{and} \quad \frac{d^2}{dx^2}(\cos x) = -\cos x,$$

conditions (V*) and (A*) are also satisfied by

$$\mathbf{r(t)} = R(\cos t)\mathbf{i} + R(\sin t)\mathbf{j},$$

with the constant K equal to R and the constant M equal to 1. Thus, this position function is a model for uniform circular motion on a circle of radius R provided that the speed of the object is also R and provided that the acceleration vector and the position vector have equal magnitudes. ∎

A good mathematical model of uniform circular motion on a circle of radius R should not require the speed of the object to be numerically equal to the radius of the path. That might be the case for a particular motion, but we need a model with more flexibility—one that allows us to specify the speed of the object as well as the radius of its circular path. That can be accomplished by introducing a parameter k into our model, as in the following generalization of the preceding theorem and proof. The proof is left to you as Problem 1.

Theorem 9.12(b) Suppose that k is a nonzero real number. Then the vector function

$$\mathbf{r_k(t)} = R\cos(kt)\mathbf{i} + R\sin(kt)\mathbf{j}$$

describes uniform circular motion for an object P on a circle of radius R centered at the origin O with a constant speed kR and a constant magnitude of acceleration k^2R.

Applying the models for uniform circular motion

We now return to one of the problems that motivated this discussion.

EXAMPLE 1 Determine the speed that must be attained by a launch vehicle to place a communications satellite in a circular orbit above the equator at a distance of 5000 miles above the center of Earth.

Solution Assume that Earth is a sphere of radius 3960 miles and that the satellite's orbit is near enough to the earth so that the acceleration due to Earth's gravity is essentially the same as that on Earth's surface, namely,

$$g = 32.2\frac{\text{ft}}{\text{sec}^2} \approx 79000\frac{\text{mi}}{\text{hr}^2}.$$

Also assume that other forces acting on the satellite, such as atmospheric drag and the gravitational attraction of the moon and other celestial bodies, are negligible compared to the gravitational attraction of the earth. (These assumptions limit the radius R of the orbit to the range from roughly 4100 to about 20,000 miles.)

By Theorem 9.12(b), we can model the uniform circular motion of the satellite so that its position at time t is given by

$$\mathbf{r(t)} = 5000\cos(kt)\mathbf{i} + 5000\sin(kt)\mathbf{j}.$$

The velocity $\mathbf{v(t)}$ of the satellite at time t is the derivative of $r(t)$.

$$\mathbf{v(t)} = -5000k\sin(kt)\mathbf{i} + 5000k\cos(kt)\mathbf{j}$$

Its speed $s(t)$ at time t is the magnitude of the velocity vector, so $s(t) = 5000k$, a constant. We call the constant speed s. Its acceleration at time t, $a(t)$ is the derivative of velocity.

$$\mathbf{a(t)} = -5000k^2\cos(kt)\mathbf{i} - 5000k^2\sin(kt)\mathbf{j}$$

The acceleration has magnitude $|\mathbf{a(t)}| = 5000k^2 = \dfrac{s^2}{5000}$.

Since the satellite's acceleration must equal the acceleration due to gravity g,

$$\frac{s^2}{5000} \approx 79000\frac{\text{mi}}{\text{hr}^2},$$

from which $s \approx \sqrt{5000 \cdot 79000} \approx 19{,}900\frac{\text{mi}}{\text{hr}}$, and $k \approx 3.98\frac{1}{\text{hr}}$.

The period p of the orbit is the period of the position vector, so $p = \frac{2\pi}{k} \approx$ 1.58 hours.

Are there other mathematical models for uniform circular motion on a circle of radius R that differ significantly from the model given in Theorem 9.12(b)? The answer is no. Under reasonable differentiability restrictions on the pair of functions g and h, the converse of Theorem 9.12(a) can be proved. That is, if g and h satisfy the three conditions

$$g(t)^2 + h(t)^2 = 1, \quad \frac{d}{dt}g(t) = -h(t), \quad \text{and} \quad \frac{d}{dt}h(t) = g(t),$$

then $g(t) = \cos t$ and $h(t) = \sin t$. Thus, not only are the sine and cosine functions useful for modeling uniform circular motion, but the mathematical model of uniform circular motion also explains why the sine and cosine functions have the first and second derivatives that they do.

9.3.3 Problems

1. Prove Theorem 9.12(b).

2. Explain why the model in Theorem 9.12(b) has the following property: Once the radius R and the speed s of the motion are specified, the constant k and the magnitude of the acceleration are determined.

3. a. Use a geometric argument based on Figure 40 to verify the limit formula

$$\lim_{\theta\to 0}\frac{\sin\theta}{\theta} = 1, \text{ where } \theta \text{ is in radians.}$$

b. Apply the limit definition of the derivative

$$\frac{df(x)}{dx} = \lim_{h\to 0}\frac{f(x+h) - f(x)}{h}$$

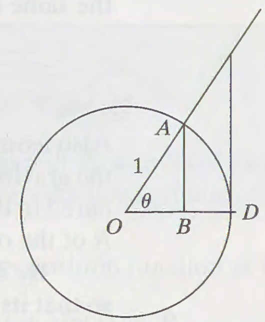
to derive the formula $\frac{d}{dx}(\sin x) = \cos x$, where x is in radians.

4. Verify that $\mathbf{r}(t) = g(t)\mathbf{i} + h(t)\mathbf{j}$ and $\mathbf{v}(t) = g'(t)\mathbf{i} + h'(t)\mathbf{j}$ are orthogonal vectors if and only if the following condition is satisfied:

$$g(t)g'(t) + h(t)h'(t) = 0 \qquad \text{for all } t.$$

5. Compute the orbital insertion speed and altitude to establish a stationary orbit above the equator for a communications satellite.

Figure 40



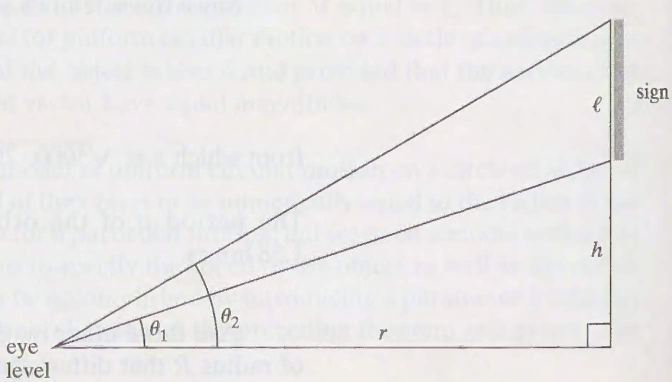
Chapter Projects

1. **Trigonometry and musical instruments.** Read "The Mathematics of Musical Instruments", by Rachel W. Hall and Kresimir Josic, *The American Mathematical Monthly* 108: 347–357 (April 2001). Write a summary of this article that could be used for presentation to a class.

2. **Regiomontanus's problem.** Imagine you are looking at a vertical sign that is above your eye level, as shown in Figure 41.

Assume that the vertical dimension ℓ of the sign and the height h of its bottom above eye level are known. Let r be the distance of your eye from the projection of the sign on the ground. We want to know the best place from which to read the sign. For best viewing of the sign we would want to make the measure of the angle $\theta_2 - \theta_1$ as large as possible. The sign then

Figure 41

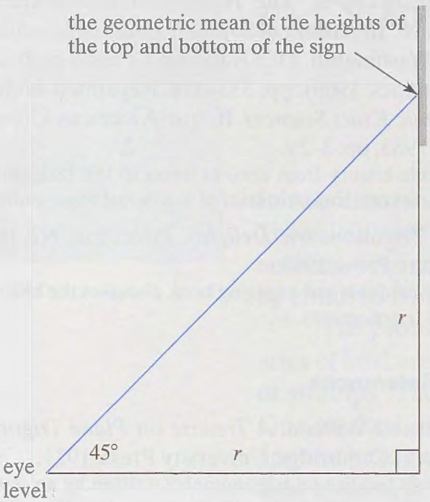


occupies the greatest portion of your field of vision. Consequently, this problem can be formulated as follows: Given fixed values of ℓ and h, find the value of r that maximizes $\theta_2 - \theta_1$.

a. Show that the maximum viewing angle occurs when the distance $r = \sqrt{h(h + \ell)}$.

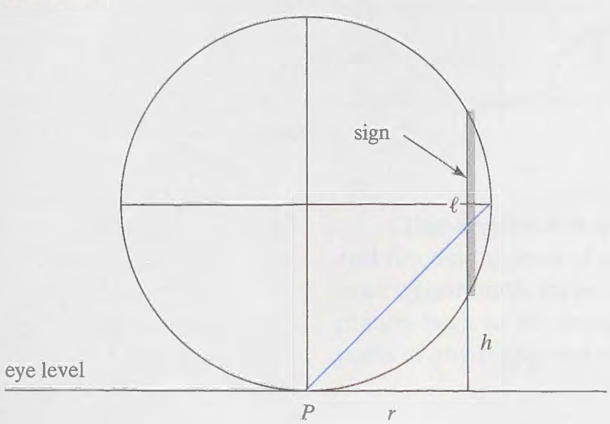
b. Interpret the result in part **a** by showing that the maximum viewing angle occurs when a line of sight from the eye elevated at a 45° angle points at the "middle" of the sign, where here the "middle" is the geometric mean of the height of the bottom of the sign and the height of the top of the sign above eye level, as shown in Figure 42.

Figure 42



c. Interpret the result in another way by showing that the maximum viewing angle occurs when the viewing point P is the point of tangency with eye level of a circle in a plane perpendicular to the ground and to the sign, tangent to the ground, and which contains the top and bottom of the sign (see Figure 43).

Figure 43



3. Constructing a table of sines and cosines. Assume you have a calculator that has no keys for the trignometric functions but can calculate sums, products, differences, quotients, and square roots. Construct an approximate table of sines and cosines for angle measures from 0° to 90° in increments of 1° with the following steps.

a. From the known values of these functions for 30° and 45°, using the identities in Section 9.3.1, obtain exact values for $\sin 15n°$ and $\cos 15n°$ for $n = 1, 4,$ and 5. Then record decimal approximations to these values.

b. Prove that $\sin 18° = \frac{\sqrt{5}-1}{4}$.

c. Use the values from parts **a** and **b** and the identities in Section 9.3.1 and its problems to obtain decimal approximations to $\sin 3n°$ and $\cos 3n°$ for $1 \le n \le 29$.

d. Interpolate from the values you found in part **c** to obtain decimal approximations for all the other integer degree measures from 0° to 90°.

e. Compare the values you obtained in part **d** with the values given by a calculator. Where you are farthest off, use the half-angle formulas and more interpolation to obtain better estimates.

4. Remarkable equalities. Values of the trigonometric functions are related to each other in many and wondrous ways. Here are eight identities and four relationships among specific values. Deduce as many as you can.

a. For any x, $\sin x + \sin(2x) + \sin(3x) + \cdots + \sin(nx) = \dfrac{\sin\frac{1}{2}(n+1)x \cdot \sin\frac{1}{2}(nx)}{\sin\frac{1}{2}x}$.

b. For any x, $\cos x + \cos(2x) + \cos(3x) + \cdots + \cos(nx) = \dfrac{\cos\frac{1}{2}(n+1)x \cdot \sin\frac{1}{2}(nx)}{\sin\frac{1}{2}x}$.

c. For any x, $\sin x + \sin(3x) + \sin(5x) + \cdots + \sin((2n-1)x) = \frac{\sin^2(nx)}{\sin x}$.

d. For any x, $\cos x + \cos(3x) + \cos(5x) + \cdots + \cos((2n-1)x) = \frac{\sin(2nx)}{2\sin x}$.

In identities **e**–**h**, A, B, and C are angles in any triangle ABC.

e. $\sin A + \sin B + \sin C = 4\cos\frac{1}{2}A \cdot \cos\frac{1}{2}B \cdot \cos\frac{1}{2}C$

f. $\cos A + \cos B + \cos C = 1 + 4\sin\frac{1}{2}A \cdot \sin\frac{1}{2}B \cdot \sin\frac{1}{2}C$

g. $\tan A + \tan B + \tan C = \tan A \cdot \tan B \cdot \tan C$

h. $\cot\frac{1}{2}A + \cot\frac{1}{2}B + \cot\frac{1}{2}C = \cot\frac{1}{2}A \cdot \cot\frac{1}{2}B \cdot \cot\frac{1}{2}C$

i. $\sin 20° \cdot \sin 40° \cdot \sin 60° \cdot \sin 80° = \frac{3}{16}$

j. $\cos 20° \cdot \cos 40° \cdot \cos 60° \cdot \cos 80° = \frac{1}{16}$

k. $\sin 6° \cdot \sin 42° \cdot \sin 66° \cdot \sin 78° = \frac{1}{16}$

l. $\sin 12° \cdot \sin 24° \cdot \sin 48° \cdot \sin 96° = \frac{1}{16}$

5. Modeling with trigonometric functions. In Section 9.2.2, the length of a day in the year 2000 at 40°N latitude over a year is approximated by a trigonometric function. Find data for your location for at least one day in each month of the current year for each of these three times: sunrise, sunset, and the length of a day. (*Note*: For sunrise and sunset, you may need to interpolate from values given in an almanac to take into account the longitude of your location.) Model each of these times with a trigonometric function and compare the values given by your model with the actual data for your location.

Bibliography

Unit 9.1 References

Aaboe, A. *Episodes from the Early History of Mathematics*. New Mathematical Library Volume 13. Washington, DC: Mathematical Association of America, 1964.
This volume includes a detailed exposition on how Ptolemy constructed a trigonometric table in his *Almagest*.

Berggren, John Lennart. *Episodes in the Mathematics of Medieval Islam*. New York: Springer-Verlag, 1986.
Chapter 5 discusses tables of chords and the sine, introduction of the six trigonometric functions, proofs of the addition formulas, among other things. Chapter 6 discusses spherical trigonometry.

Gelfand, I. M., and Mark Saul. *Trigonometry*. Boston: Birkhäuser, 2001.
This short text treats in a simple and elegant fashion all of the trigonometric results usually seen in high school, and many others as well, often with interesting historical information added. Calculus is not used, but the discussion is much broader and fuller than is typical in most precalculus texts.

Unit 9.2 References

Dunham, William. *Euler: The Master of Us All*. Washington, DC: Mathematical Association of America, 1999.
Euler's work with complex numbers and trigonometry is found in Chapter 5.

Katz, Victor. *A History of Mathematics*. New York: HarperCollins, 1993.
Attention to the history of trigonometry, including non-European work, is found in Chapter 4 and Sections 6.6, 7.4, and 10.3.

Kennedy, Edward S. "The History of Trigonometry—An Overview." In *Historical Topics for the Mathematics Classroom* (Washington, DC: National Council of Teachers of Mathematics, 1969), pp. 333–359. Reprinted in *Studies in the Islamic Exact Sciences*. Beirut: American University of Beirut, 1983, pp. 3–29.
A readable history from ancient times to the 18th century, followed by several short histories of individual trigonometric ideas.

Maor, Eli. *Trigonometric Delights*. Princeton, NJ: Princeton University Press, 1998.
This very readable and engaging book discusses the historical evolution of trigonometry.

Unit 9.3 References

Hobson, Ernest William. *A Treatise on Plane Trigonometry*. New York: Cambridge University Press, 1921.
This classic treatise on trigonometry written by an outstanding mathematician of that time demonstrates that the subject has a richness and depth that is not at all evident in most modern textbooks.

Chapter

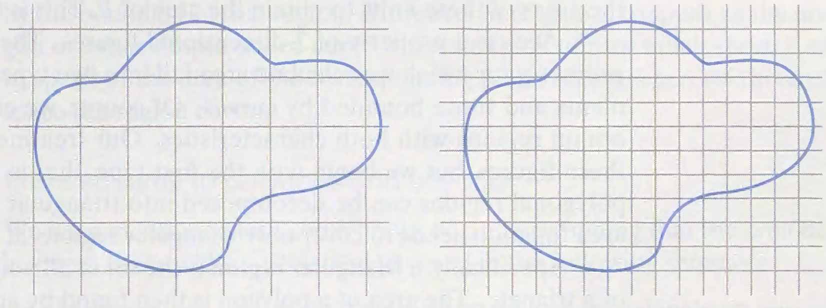
10

AREA AND VOLUME

In two-dimensional space, area measures the space covered or enclosed by a figure. The calculation of area of land is one of the oldest problems that we know of in mathematics, originating in part from problems of fair apportionment of land inherited by children from their parents.

The problem of calculating such areas immediately becomes difficult. Boundaries of land are often not straight; they may include parts of rivers or edges of foothills or swamps. The land itself may include small bodies of water. The land may be hilly. One method of finding the area of irregular pieces of land is not very far from the approach taken in calculus. If the land is not too large, it is photographed from above, thus making it flat. A grid of congruent squares with known dimensions is placed over the photo (Figure 1a). The number of squares entirely inside the boundary plus half the number crossed by the boundary provides an estimate for the area. If a more precise estimate is desired, then smaller squares are used (Figure 1b).

Figure 1



(a) (b)

This approach is quite similar to the approach we take in Unit 10.1 to develop and discuss the area of plane figures. It applies also to the calculation of the surface area of common three-dimensional figures. That is, we can trace everything ultimately back to the areas of squares because, as we will show, we can rearrange the parts of any polygonal region to form a square.

In Unit 10.2 we discuss volume. The volume of a 3-dimensional figure is a measure of the space occupied or enclosed by it. The approach that is used for area would suggest that the volume of a polyhedron can be determined by rearranging its parts to form a cube. It turns out that, although we can and do measure volume in cubic units, we cannot reduce all three-dimensional polyhedra to cubes the way that we can reduce all two-dimensional polygons to squares. Nonetheless, we can still obtain formulas for the volume of the three-dimensional figures commonly studied in school.

Thus, despite important similarities, area and volume possess significant differences that make them interesting to study and compare.

Unit 10.1 Area

Many students leave their study of mathematics associating area with formulas (rather than size) and thinking that they need to have the formula for the area of a figure before they can obtain its area. In this unit, we show many ways of arriving at the area of a figure, and many ways in which area formulas are related. We begin at the logical beginning, with a definition of area.

In comparison with distance as linear measure, area can be conceived of as surface measure—one that gives the plane or surface "content" of a figure. And, like distance, the calculation of area is subject to the context in which it is applied. In geometry we calculate area using formulas based on the sides and angles of figures. In calculus, we define area as the limit of a sum and calculate it as a definite integral. Where did these methods come from? What assumptions do we make when we apply them? These are some of the questions we address in this unit.

10.1.1 What is area?

Suppose we want to calculate the area of a plane region F and we do not know any formulas for calculating the area of F. How could we proceed? One possible method is to compare F with a region whose area we do know. If we can superimpose that region on F, we can be confident both have the same area.

Another method we could use is to choose a region and designate it as "the unit of area". Then we calculate the area of F by decomposing it into nonoverlapping units that completely cover the region. This decomposition method presupposes that we can decompose F into a finite number of nonoverlapping units. We then add the areas of these units to obtain the area of F. This is the approach we follow.

Area is a property of 2-dimensional figures. The basic regions of elementary geometry for which we calculate area fall into two types: those bounded by line segments and those bounded by curves. Of course, we can combine these regions to obtain regions with both characteristics. Our treatment of area will extend to all these figures, but we begin with the first type, that is, polygonal regions. Because polygonal regions can be decomposed into triangular regions, our definition of an area function needs to cover only triangular regions at first.

Specifically, a **triangular region** is the set of all points interior to or on the sides of a triangle. The area of a polygon is then found by adding up the areas of its constituent triangular regions. Before we do this, however, we need to choose a unit of area. We choose the square as the unit of area for several reasons. It has congruent edges, so its area calculation is the same no matter which edge you choose, unlike the triangle whose area calculation depends on what side is its base. Also, it turns out (as we show in Section 10.1.5) that any union of triangular regions can be decomposed and rearranged into a single square. And because we customarily measure in *square units*, we can transfer the theory based on the square as a unit of area directly to practical problems.

Definition Let F be the union of a finite collection of triangular regions in E^2. An **area function** α is a function that assigns to each such F a positive real number $\alpha(F)$ such that:

1. If $F_1 \cong F_2$, then
$$\alpha(F_1) = \alpha(F_2). \qquad \text{(Congruence property)}$$
2. If the triangular regions making up F_1 and F_2 have no interior points in common, then
$$\alpha(F_1 \cup F_2) = \alpha(F_1) + \alpha(F_2). \qquad \text{(Additive property)}$$
3. If F is a square region with side x, then
$$\alpha(F) = x^2. \qquad \text{(Area of square)}$$

Some books replace (3) by the area $\alpha(F) = ab$ of a rectangular region with adjacent sides a and b. We could also replace (3) by the area $\alpha(F) = \frac{1}{2}x^2$ of an isosceles right triangle with leg x (i.e., half a square), which puts the entire definition in terms of triangular regions. We begin with the area of a square because area is typically measured in square units and because we can derive these other formulas quite easily from it, as you will see.

Part (2) of this definition rests on a tacit assumption: that if a region is split into triangular regions in two different ways, then both the areas calculated are the same. That assumption relies on a further assumption, that the pieces of a triangular region can be rearranged into a square, and if there is more than one rearrangement into squares, then all such squares have the same size.[1] You might think of this in the following way: Suppose you calculate the area of a triangle using the familiar formula $A = \frac{1}{2}bh$, that is, area = half the product of a length of a side and the altitude to that side. Will you get the same area if you pick another side of the same triangle? The answer is yes, but this is not something that a complete mathematical theory could take for granted. In Section 10.1.2, we return to this question.

The theory of area technically requires that we always refer to the areas of regions because area is defined in terms of unions of sets. But in practice, we speak of the area of a square or triangle, not the area of a square region or triangular region. There is nothing mathematically wrong with this. We are merely referring to a region by the boundary of that region. Moreover, it is natural to speak of the area of a polygon or other simple closed curve when we think of how much space it encloses, and to speak of the area of the corresponding plane region when we think of how much space the region covers.

From squares to polygons and beyond

We have mentioned that some books begin by assuming that the area of a rectangle is length times width. They then proceed in the following sequence:

area of rectangle ⇒ area of right triangle ⇒ area of any triangle ⇒ area of trapezoid.

In contrast, we begin with the more basic area formula for a square and go as far as area calculations using calculus. Along the way we stop to gaze at the landscape and deduce some broadly applicable theorems not found in all high school books. Our development is described schematically in Table 1.

[1] For a detailed treatment, see *Geometry: A Metric Approach with Models*, by Richard S. Millman and George D. Parker (New York: Springer-Verlag, 1981).

Table 1

Area of square (def. of area function)	⇒	Area of rectangle (Theorem 10.1)	⇒	Riemann sums (Section 10.1.4)		
		⇓				
		Area of right triangle (Corollary)	⇒	SASΔ formula (Theorem 10.8)	⇒	ASAΔ formula (Theorem 10.9)
		⇓				
Area of quadrilateral with $\perp$ diagonals (Theorem 10.4)	⇐	Area of triangle (Theorem 10.3)	⇒	Area of circumscribed polygon (Theorem 10.6)	⇒	Hero's formula (Theorem 10.7)
		⇓		⇓		
		Area of trapezoid (Theorem 10.5)		Area of circle (Theorem 10.10)		
		⇓		⇓		
		Trapezoidal rule		Area of ellipse		

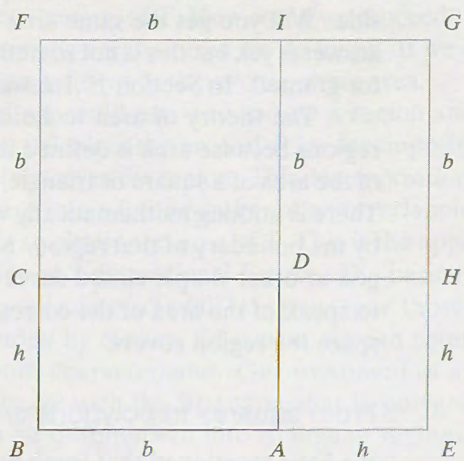
From squares to rectangles

The proof of Theorem 10.1 shows how the formula for the area of a rectangle follows from the definition of an area function. The proof is short and involves only elementary algebra, yet it uses all three parts of the definition.

Theorem 10.1 The area of a rectangle is the product of the lengths of two adjacent sides.

Proof: Without loss of generality, let $ABCD$ be the rectangle and let $AB = b$ and $BC = h$. Extend $\overline{BA}$ by a line segment of length h to the point E, and extend $\overline{BC}$ by a segment of length b to the point F. $\overline{BE}$ and $\overline{BF}$ have length $(h + b)$ and are the sides of a square $EBFG$, as shown in Figure 2.

Figure 2



Rectangles $ABCD$ and $DHGI$ are congruent. By the congruence property (1) of the definition of area function, they have the same area. By property (3) of the definition of area function, $\alpha(EBFG) = (h + b)^2$, $\alpha(EADH) = h^2$, and $\alpha(\text{DCFI}) = b^2$. ■

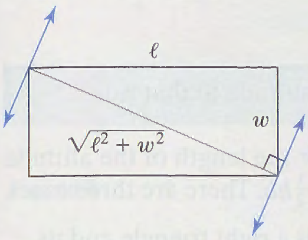
Question: Complete the proof, indicating where the additive property (2) of the area function is used.

In the statement of Theorem 10.1, the word "side" refers to a segment. But often we use that word to refer to the segment's length. For instance, we speak of a rectangle with sides a and b. The dual use of a word to refer either to a segment or its length is found throughout geometry, for example, with "radius", "diameter", "leg", "hypotenuse", "diagonal", etc.

Since a rectangle with dimensions a and b can be formed from two congruent right triangles with legs a and b, a formula for the area of a right triangle follows immediately from Theorem 10.1.

Corollary: The area of a right triangle is $\frac{1}{2}$ the product of its legs.

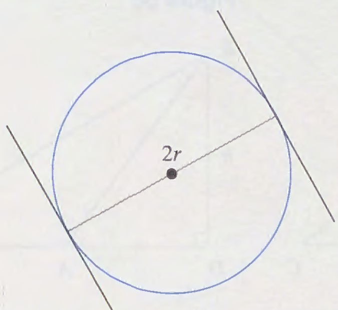
Figure 3



Other measures of two-dimensional regions

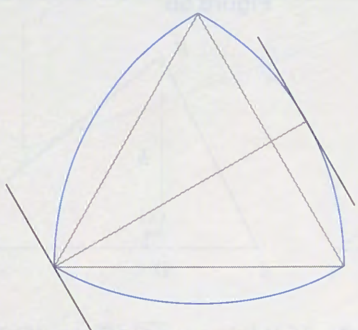
Area is not the only measure of two-dimensional regions. For example, the **width** of a two-dimensional region in a particular direction is the length of the longest segment joining two points of the region parallel to that direction. For instance, since the longest segments joining two points on a rectangle are its diagonals, the longest width of a rectangle with dimensions ℓ and w is $\sqrt{\ell^2 + w^2}$, in the direction of either diagonal (Figure 3). A circle with radius r has constant width $2r$ and is one of many figures with constant width. Three such figures are displayed in Figures 4a–c.

Figure 4a



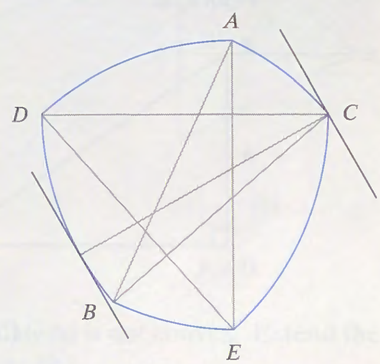
circle

Figure 4b



Reuleaux Triangle

Figure 4c



"pentagon" of constant width

And of course there is *perimeter*. The **perimeter** $p(R)$ of any polygonal region R is the sum of the lengths of its sides. This defines a function p that students often confuse with area. One reason for the confusion is that the units of the function are ignored in the mathematical definition, so that the ranges of the area and perimeter functions are identical. However, the units are never the same: The unit of perimeter is a 1-dimensional unit segment and the unit of area is the 2-dimensional unit square. A second reason for the confusion is that the numbers used for dimensions of rectangles and triangles in schoolbooks are often small integers, and for these integers the numerical values of area and perimeter are of the same general size. A third reason for the confusion may be that these topics are taught without reference to physical examples. With examples, the difference is significant: A lake's area indicates how much room there is for fishing, while its perimeter gives the amount of shoreline; a room's area tells how much flooring is needed, while its perimeter gives the amount of baseboard; and so on.

The following theorem establishes that area does not determine perimeter even in simple figures. The figure of Theorem 10.1 suggests a geometric proof for Theorem 10.2 that we encourage you to produce. Here we provide an elegant algebraic proof. The key to the proof of Theorem 10.2 is to let x be the amount by which a side differs from $\frac{1}{4}$ the perimeter p. When the rectangle is a square, $\frac{1}{4}p$ is the length of the each side.

Theorem 10.2 Of all rectangles with a given perimeter, the square has the greatest area.

Proof: Let the rectangle have perimeter p, length ℓ, width w, and area A. By the definition of perimeter, $p = 2\ell + 2w$, so $\frac{p}{2} = \ell + w$. We know then that ℓ is some number between 0 and $\frac{p}{2}$. So we can let $\ell = \frac{p}{4} + x$. Then $w = \frac{p}{4} - x$. So $A = \ell w = \left(\frac{p}{4} + x\right)\left(\frac{p}{4} - x\right) = \frac{p^2}{16} - x^2$. Since x^2 is always nonnegative, the area A is maximized when $x = 0$: that is, when $\ell = w$ and the rectangle is a square. ┘

Area formulas quickly deduced

From the area formula for a right triangle (Corollary to Theorem 10.1), we can deduce the familiar area formula for any triangle.

Theorem 10.3 The area of a triangle is $\frac{1}{2}$ the product of a side and the altitude to that side.

Proof: Let the triangle be ABC, let $b = AC$, and let $h =$ the length of the altitude $\overline{BD}$ from B to $\overline{AC}$. We wish to show that $\alpha(\Delta ABC) = \frac{1}{2}hb$. There are three cases.

Case 1: If $D = A$ or $D = C$ (Figure 5a), then ΔABC is a right triangle and its area is $\frac{1}{2}hb$ from the Corollary to Theorem 10.1.

Figure 5a

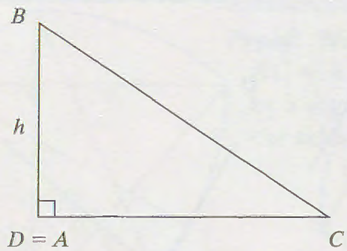


Figure 5b

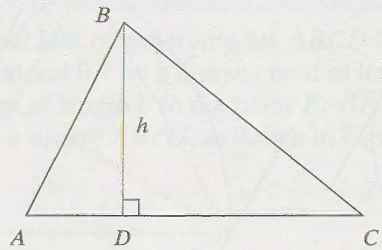
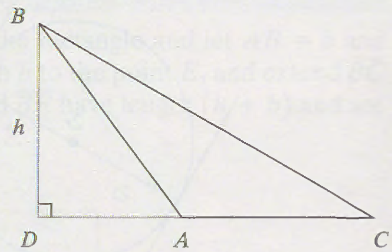


Figure 5c



Case 2: If D lies on $\overline{AC}$ (Figure 5b), then, using the Corollary to Theorem 10.1.

$$\alpha(\Delta ABC) = \frac{1}{2}h \cdot AD + \frac{1}{2}h \cdot DC \qquad \text{(from the additive property of the area function)}$$

$$= \frac{1}{2}h(AD + DC) = \frac{1}{2}hb.$$

Case 3: If D lies outside of $\overline{AC}$, say on $\overrightarrow{CA}$ as shown (Figure 5c), then, again from the additive property of the area function,

$$\alpha(\Delta ABC) + \frac{1}{2}h \cdot AD = \frac{1}{2}h \cdot DC.$$

Thus

$$\alpha(\Delta ABC) = \frac{1}{2}h \cdot DC - \frac{1}{2}h \cdot AD$$

$$= \frac{1}{2}h(DC - AD)$$

$$= \frac{1}{2}hb.$$

If D lies outside of $\overline{AC}$ on $\overrightarrow{AC}$, the proof is similar to that in (3). ┘

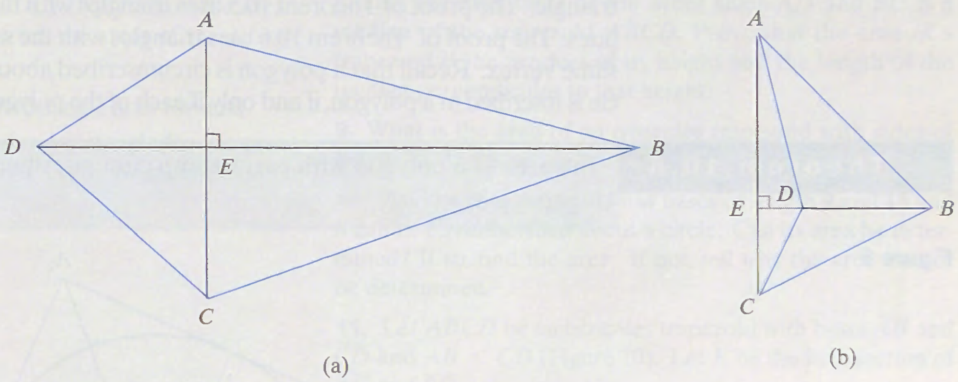
From either the formula for the area of a rectangle or for the area of a triangle, a formula for the area of any quadrilateral with perpendicular diagonals can be obtained. This formula is useful because it applies to all kites and thus to all rhombi and squares.

Theorem 10.4 If a quadrilateral has perpendicular diagonals of lengths d_1 and d_2, then its area is $\frac{1}{2}d_1 d_2$.

Proof: Let $ABCD$ be a quadrilateral with perpendicular diagonals $\overline{AC}$ and $\overline{BD}$ intersecting at E, as shown in Figure 6a. From the additive property of the area function, and from Theorem 10.3,

$$\alpha(ABCD) = \alpha(\Delta ADB) + \alpha(\Delta CDB) = \frac{1}{2}\cdot AE\cdot BD + \frac{1}{2}\cdot EC\cdot BD = \frac{1}{2}\cdot AC\cdot BD.$$

Figure 6



If the diagonals do not intersect, then the quadrilateral is not convex. Extend the diagonals as needed to again intersect at E (Figure 6b).

Now

$$\alpha(\Delta ABC) = \alpha(ABCD) + \alpha(\Delta ADC).$$

So

$$\frac{1}{2}\cdot AC\cdot BE = \alpha(ABCD) + \frac{1}{2}\cdot AC\cdot DE$$

$$\frac{1}{2}\cdot AC\cdot BE - \frac{1}{2}\cdot AC\cdot DE = \alpha(ABCD)$$

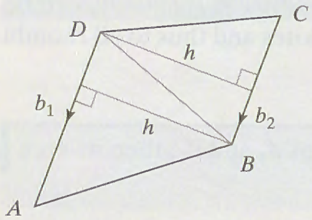
$$\frac{1}{2}\cdot AC\cdot BD = \alpha(ABCD).$$

┘

From the formula for the area of a triangle, we can obtain a formula for the area of any trapezoid. By the shortness of its proof, it could be considered a corollary to Theorem 10.3. Its importance leads us to call it a theorem.

Theorem 10.5 The area of a trapezoid with bases b_1 and b_2 and height h is $\frac{1}{2}h(b_1 + b_2)$.

Figure 7



Proof: Refer to Figure 7. Suppose $\overline{AD}$ and $\overline{BC}$ are bases of trapezoid $ABCD$, h is the distance between $\overline{AD}$ and $\overline{BC}$, $b_1 = AD$, and $b_2 = BC$. Then

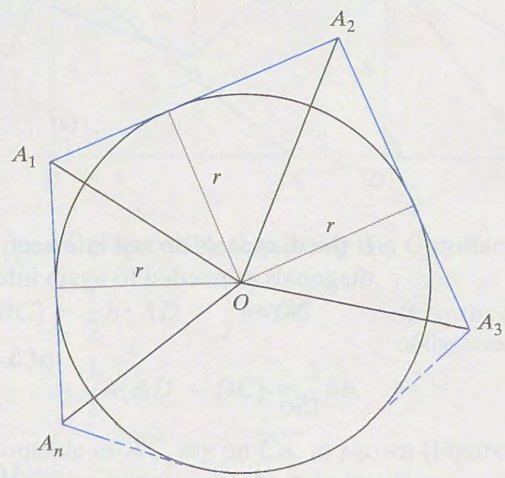
$$\begin{aligned}\alpha(ABCD) &= \alpha(\Delta ABD) + \alpha(\Delta DBC) &&\text{by property (2) of the area function}\\ &= \frac{1}{2}b_1h + \frac{1}{2}b_2h\\ &= \frac{1}{2}h(b_1 + b_2).\end{aligned}$$

The trapezoid area formula is significant because it applies to all parallelograms, rhombi, rectangles, and squares, and in calculus is used to develop the Trapezoidal Rule that approximates areas under curves. (In calculus, the bases are typically vertical, so some students at first do not recognize that the figures are trapezoids.)

In Section 8.3.2, in two ways we extended a theorem about a line intersecting a side of a triangle. One way was by increasing the number of parallel lines; the second was by increasing the number of rays from one vertex. Theorem 10.5 and the next theorem roughly arise applying the same types of generalization to the area formula $A = \frac{1}{2}hb$ for a triangle. The proof of Theorem 10.5 uses triangles with the same height between parallel lines. The proof of Theorem 10.6 uses triangles with the same height emanating from the same vertex. Recall that a polygon is circumscribed about a circle, or, equivalently, a circle is inscribed in a polygon, if and only if each of the polygon's sides is tangent to the circle.

Theorem 10.6 The area of a polygon with perimeter p circumscribed about a circle with radius r is $\frac{1}{2}rp$.

Figure 8



Proof: The proof is straightforward. Let $A_1A_2A_3 \ldots A_n$ be a polygon circumscribed about circle O with radius r (see Figure 8). The polygonal region $A_1A_2A_3 \ldots A_n$ is the union of the triangular regions $OA_1A_2, OA_2A_3, \ldots, OA_nA_1$. Each triangular region has height r because a radius of a circle is perpendicular to any tangent to the circle at the point of tangency. So

$$\begin{aligned}\alpha(A_1A_2A_3 \ldots A_n) &= \alpha(OA_1A_2) + \alpha(OA_2A_3) + \cdots + \alpha(OA_nA_1)\\ &= \frac{1}{2}r(A_1A_2) + \frac{1}{2}r(A_2A_3) + \cdots + \frac{1}{2}r(A_nA_1)\\ &= \frac{1}{2}r(A_1A_2 + A_2A_3 + \cdots + A_nA_1)\\ &= \frac{1}{2}rp.\end{aligned}$$

Theorem 10.6 applies to all regular polygons, and also to any triangle, since a circle can be inscribed in any of these figures. The application to triangles, discussed in the next section, is particularly productive.

10.1.1 Problems

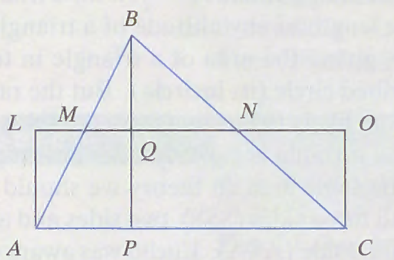
1. a. Estimate the area of the region in Figure 1a given that each square is 1 unit on a side.

b. Estimate the area of the region in Figure 1b given that each square is $\frac{1}{2}$ unit on a side.

2. In the proof of Theorem 10.1, if the original rectangle $ABCD$ has dimensions 7″ and 10″, what are the dimensions of the two squares $HEAD$ and $FIDC$? Explain how the area of the rectangle can be derived from the areas of these two squares.

3. Figure 9 suggests a proof of the triangle area formula $A = \frac{1}{2}bh$ from the area formula for a rectangle for the case where the angles including the base of the triangle are both acute.

Figure 9



a. Provide the proof.

b. Draw the corresponding figure for the case in which a base angle is obtuse and prove the theorem for this case.

4. A rectangle has dimensions x and y, with $x \geq y$.

a. What is its minimal width?

b. What is its maximal width?

c. If this rectangle is the frame of a door, what is the radius of the largest circular table top that can be passed through the door? (Ignore the thickness of the table top.)

5. Let ℓ and m be parallel lines containing the vertices A and C of the rectangle $ABCD$ and not containing any other points of the rectangle. Prove that the distance between ℓ and m is greatest when either line is perpendicular to $\overline{AC}$.

6. Let ΔABC be equilateral. Consider the region bounded by the circular arc $\widehat{BC}$ with center A, the circular arc $\widehat{AC}$ with center B, and the circular arc $\widehat{AB}$ with center C. Prove that this region has constant width. (The union of the three arcs $\widehat{BC}$, $\widehat{AC}$, and $\widehat{AB}$ is known as a **Reuleaux triangle** and is shown in Figure 4b of this section.)

7. Provide a proof for the situation of Figure 6a of Theorem 10.4 that avoids Theorem 10.3 by separating the quadrilateral into four right triangles.

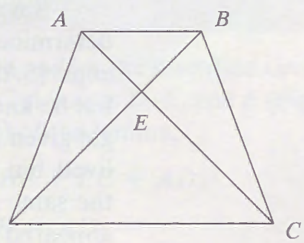
8. Let $ABCD$ be a trapezoid with $\overline{AB} \parallel \overline{CD}$. The segment joining the midpoints of the other sides $\overline{AD}$ and $\overline{BC}$ is a **median** of the trapezoid $ABCD$. Prove that the area of a trapezoid is the product of its height and the length of the median perpendicular to that height.

9. What is the area of an isosceles trapezoid with sides of length 20, 25, 25, and 28?

10. An isosceles trapezoid has bases of length 9 and 15 and it can be circumscribed about a circle. Can its area be determined? If so, find the area. If not, tell why the area cannot be determined.

11. Let $ABCD$ be an isosceles trapezoid with bases $\overline{AB}$ and $\overline{CD}$ and $AB < CD$ (Figure 10). Let E be the intersection of $\overline{AC}$ and $\overline{BD}$.

Figure 10



a. Prove: $\alpha(\Delta BEC) = \alpha(\Delta AED)$.

b. Prove: $\alpha(\Delta ABE) < \alpha(\Delta EDC)$.

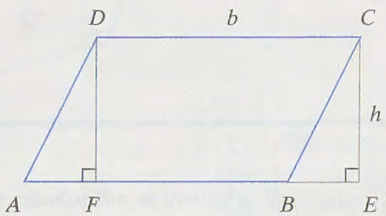
12. A rhombus has diagonals of lengths a and b. Show three different ways of finding its area.

13. Let $ABCD$ be a parallelogram and let E be a point on $\overline{AC}$. Let the parallel to $\overline{AD}$ through E intersect $\overline{AB}$ at H and $\overline{CD}$ at I. Let the parallel to $\overline{AB}$ through E intersect $\overline{AD}$ at F and $\overline{BC}$ at G. Prove that parallelograms $HEGB$ and $FDIE$ have equal area. (This is a theorem that Euclid exploited in his development of area.)

14. Let $E, F, G,$ and H be the midpoints of the sides $\overline{AB}, \overline{BC}, \overline{CD},$ and $\overline{DA}$ of convex quadrilateral $ABCD$. Prove that $\alpha(EFGH) = \frac{1}{2}\alpha(ABCD)$.

15. a. An area formula for a parallelogram, $A = hb$, is often found by finding a rectangle with equal area, as in Figure 11a. Explain this argument.

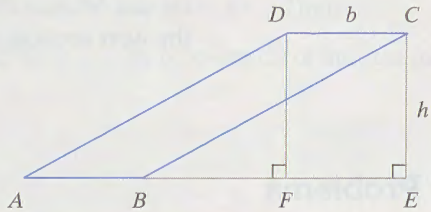
Figure 11a



(a)

b. The argument of part **a** does not work for parallelograms like those in Figure 11b. Write a proof that $\alpha(ABCD) = hb$ for the situation in Figure 11b.

Figure 11b



(b)

16. A regular n-sided polygon is inscribed in a unit circle.

a. Prove that its perimeter is $2n \sin \frac{180°}{n}$.

b. Prove that its area is $\frac{n}{2} \sin \frac{360°}{n}$.

c. Modify the formulas of parts **a** and **b** if the circle has radius r.

17. Deduce formulas similar to those found in Problem 16 for a regular n-gon that is circumscribed about a circle.

ANSWER TO QUESTION

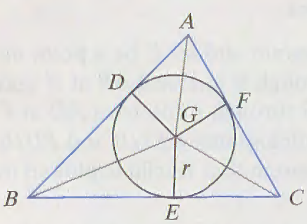
By properties (1) and (2) of the definition of area function, $\alpha(EBFG) = \alpha(EADH) + \alpha(ABCD) + \alpha(DHGI) + \alpha(FCDI) = \alpha(EADH) + 2\alpha(ABCD) + \alpha(FCDI)$. Now use property (3) and substitute $(h + b)^2 = h^2 + 2\alpha(ABCD) + b^2$. Consequently, $h^2 + 2hb + b^2 = h^2 + 2\alpha(ABCD) + b^2$. Solving the equation for $\alpha(ABCD)$, we obtain $\alpha(ABCD) = bh$.

10.1.2 Area formulas for triangles

In Section 10.1.1, we deduced the familiar area formula $A = \frac{1}{2}bh$ for a triangle. However, in practice one may not know the length of any altitude of a triangle. We also arrived at a second formula, $A = \frac{1}{2}rp$, giving the area of a triangle in terms of its perimeter p and the radius r of its inscribed circle (its **incircle**). But the radius of the incircle is also unlikely to be known. More likely to be known are the sides and angles of the triangle. So we search for an area formula using only sides and angles.

Since congruent triangles have the same area, in theory we should be able to determine the area of a triangle given all three sides (SSS), two sides and an included angle (SAS), or two angles and the included side (ASA). Euclid was aware of this fact, but he knew of no such formulas. The first of these formulas, for the area of a triangle given its sides, was discovered by Archimedes about a half century after Euclid lived, but it is known today as *Hero's Formula*, after Hero the Elder of Alexandria, the same mathematician mentioned in Section 8.1.2 in connection with optics. It appeared in Hero's book *Geodesy* and was proved in two other books, *Dioptra* and *Metrica*. The proof is quite long. To help in understanding it, we split it into two parts and call the first part a lemma.[2]

Figure 12



Lemma: Let D, E, and F be the points of tangency of the inscribed circle G of ΔABC and its sides $\overline{AB}$, $\overline{BC}$, and $\overline{AC}$, as shown in Figure 12. Let r be the radius of circle G and s be half the perimeter of ΔABC. Then

1. $\alpha(\Delta ABC) = sr$

2. $AD = AF = s - BC$
$BE = BD = s - AC$
$CF = CE = s - AB$.

[2] A **lemma** is a statement that is used in a proof of a theorem that immediately follows the lemma. The lemma is itself proved, but usually is not important enough to be labeled a theorem.

Proof: Let p be the perimeter of ΔABC. By Theorem 10.6 of Section 10.1.1,

$$\alpha(\Delta ABC) = r \cdot \frac{1}{2}p = rs.$$

Then

$$p = 2s = AD + DB + BE + EC + CF + FA.$$

$AD = AF$ because they are tangents from a point to the circle. ($\Delta ADG \cong \Delta AFG$ by HL Congruence.) Similarly, $BE = BD$ and $CF = CE$. So

$$2s = 2AD + 2BE + 2CF.$$

Dividing both sides by 2,

$$s = AD + BE + CF.$$

Now we can deduce any one of the three results by solving this equation for the desired distance. For instance,

$$\begin{aligned} AD &= s - BE - CF \\ &= s - BE - CE \\ &= s - BC. \end{aligned}$$

The other two parts follow in the same manner. ┘

The quantity s in the lemma is half the perimeter of ΔABC, so it is known as the **semiperimeter** of ΔABC. Notice that if the sides of ΔABC are named in the traditional manner as a, b, and c, then $AD = s - a$, $BE = s - b$, and $CF = s - c$. These lengths play major roles in Hero's Formula for the area of any triangle.

Theorem 10.7 **(Hero's Formula):** The area of a triangle with sides a, b, and c and semiperimeter s is $\sqrt{s(s-a)(s-b)(s-c)}$.

Proof 1 (synthetic): Let ABC be the triangle and let the inscribed circle for the triangle have center G and be tangent to ABC at points D, E, and F (Figure 13). Let $r = GD = GE = GF$. Because of part (1) of the lemma,

(1) $$\alpha(\Delta ABC) = sr = (BE + EC + AD)r.$$

Now extend $\overline{CB}$ beyond B to H so that $BH = AD$.
Substituting into (1),

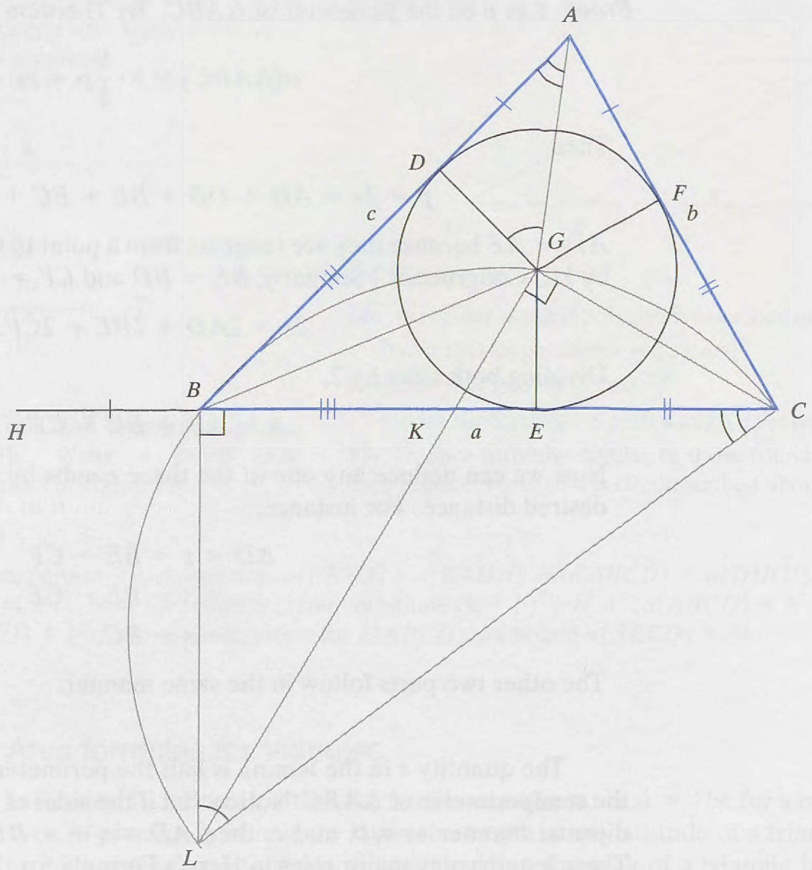
$$\begin{aligned} \alpha(\Delta ABC) &= (BE + EC + BH)r \\ &= CH \cdot r \end{aligned}$$

(2) $$\alpha(\Delta ABC) = CH \cdot EG.$$

The goal now is to get some ratios involving CH and EG, and Hero does this through similar triangles. Let L be the intersection of the perpendiculars to $\overline{CG}$ at G and $\overline{CB}$ at B. $\angle CGL$ and $\angle CBL$ are right angles, so the midpoint of $\overline{LC}$ (not drawn) is equidistant from L, B, G, and C. This means that quadrilateral $LBGC$ is inscribed in a semicircle, and so its opposite angles CGB and CLB are supplementary. But also, $\angle CGB$ and $\angle AGD$ are supplementary (because the quadrilaterals $ADGF$, $CFGE$, and $BEGD$ are kites and their symmetry diagonals bisect the central angles). Consequently, $m\angle CLB = m\angle AGD$, and since these are acute angles in right triangles,

(3) $$\Delta CBL \sim \Delta ADG.$$

Figure 13



From (3) and recalling that $BH = AD$ and $DG = EG$,

(4) $$\frac{CB}{BL} = \frac{AD}{DG} = \frac{BH}{EG}$$

From the left and right ratios in (4), $\frac{CB}{BH} = \frac{BL}{EG}$. With K being the intersection of $\overline{GL}$ and $\overline{BC}$, $\Delta GKE \sim \Delta LKB$, so that

(5) $$\frac{CB}{BH} = \frac{BL}{EG} = \frac{BK}{KE}.$$

Adding $\frac{BH}{BH}$ to the left and $\frac{KE}{KE}$ to the right ratios in (5),

(6) $$\frac{CH}{BH} = \frac{BE}{EK}.$$

With some multiplication by 1 and recognizing a geometric mean,

(7) $$\frac{CH^2}{CH \cdot BH} = \frac{BE \cdot EC}{EK \cdot EC} = \frac{BE \cdot EC}{EG^2}.$$

Again use the left and right ratios.

(8) $$\begin{aligned} CH^2 \cdot EG^2 &= CH \cdot BH \cdot BE \cdot EC \\ &= CH \cdot AD \cdot BE \cdot CF \end{aligned}$$

Now use part (2) of the lemma to rewrite the right side of (8) in terms of s, a, b, and c, and notice that $CH = s$. Rewrite the left side using (2) of this proof,

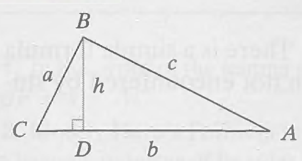
$$(9) \qquad \alpha(\Delta ABC)^2 = CH \cdot AD \cdot BE \cdot CF = s(s-a)(s-b)(s-c).$$

$$(10) \qquad \text{So} \quad \alpha(\Delta ABC) = \sqrt{s(s-a)(s-b)(s-c)}.$$

Hero's proof of the formula bearing his name is extraordinary because the result is algebraically quite complex and Hero had none of today's algebraic notation at his disposal. We have presented it to show that Hero's formula is geometrically derivable from the other area formulas, and also to exemplify the range of synthetic geometry proofs.

A shorter proof uses the Law of Cosines and quite a bit of algebraic manipulation.

Figure 14



Proof 2 (analytic): The triangle ΔABC must have at least one acute angle. Let it be $\angle A$. Let D be the foot of the altitude from B to $\overline{AC}$ (Figure 14). Let $h = BD$. We know $\alpha(\Delta ABC) = \frac{1}{2}bh$. So we will search for a formula for h in terms of the sides a, b, and c. By the Law of Cosines,

$$a^2 = b^2 + c^2 - 2bc \cos A.$$

Now $AD = c \cos A$. Consequently,

$$a^2 = b^2 + c^2 - 2b \cdot AD.$$

We solve this for AD.

$$AD = \frac{b^2 + c^2 - a^2}{2b}$$

Using the Pythagorean Theorem, we can find an expression for h.

$$h^2 = c^2 - AD^2 = c^2 - \left(\frac{b^2 + c^2 - a^2}{2b}\right)^2 = \frac{4b^2c^2 - (b^2 + c^2 - a^2)^2}{4b^2}$$

Factor the difference of squares in the numerator of the last expression, rearrange the factors as differences of squares, and factor again to obtain

$$h^2 = \frac{(2bc + b^2 + c^2 - a^2)(2bc - b^2 - c^2 + a^2)}{4b^2}$$

$$= \frac{[(b^2 + 2bc + c^2) - a^2][a^2 - (b^2 - 2bc + c^2)]}{4b^2}$$

$$= \frac{(b + c + a)(b + c - a)(a + b - c)(a - b + c)}{4b^2}.$$

Now we introduce the semiperimeter s. If $s = \frac{1}{2}(a + b + c)$, then

$$b + c + a = 2s$$
$$b + c - a = 2(s - a)$$
$$a + b - c = 2(s - c)$$
$$a - b + c = 2(s - b).$$

[These expressions are the equivalent of part (2) of the lemma before Theorem 10.7.] By substitution,

$$h^2 = \frac{[2s][2(s-a)][2(s-b)][2(s-c)]}{4b^2} = \frac{4s(s-a)(s-b)(s-c)}{b^2}.$$

So

$$h = \frac{2\sqrt{s(s-a)(s-b)(s-c)}}{b}.$$

Thus, $\alpha(\Delta ABC) = \frac{1}{2}bh = \sqrt{s(s-a)(s-b)(s-c)}$.

Both Hero's Formula and the formula $A = \frac{1}{2}rp$ for the area of a triangle are *symmetric*, in that the three sides of the triangle each play the same role. They show that the area of a triangle does not depend on a particular side being singled out as special, and that regardless of which side might be called a base, the area of the triangle would be the same. This guarantees that, once a unit has been picked, the area of a triangle is unique.

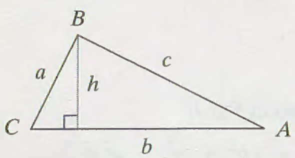
Formulas when SAS or ASA is known

Hero's Formula produces the area of a triangle given SSS. There is a simple formula for the area of a triangle given SAS. This formula is often not encountered by students because area is studied before trigonometry.

Theorem 10.8 **(SAS Area Formula):** For all triangles ABC,

$$\alpha(\Delta ABC) = \frac{1}{2}ab\sin C.$$

Figure 15



Proof: In ΔABC, let h be the altitude to side b (Figure 15). Then $\frac{h}{a} = \sin C$. Solving for h and substituting into the area formula $\alpha(\Delta ABC) = \frac{1}{2}bh$ yields this theorem.

Corollary: The area of the parallelogram with adjacent sides a and b and included angle ϕ is $ab\sin\phi$.

The SAS area formula is not symmetric. Depending on which side lengths are known, we could have any of the following variants: $\alpha(\Delta ABC) = \frac{1}{2}ab\sin C = \frac{1}{2}bc\sin A = \frac{1}{2}ac\sin B$. We use these variants to develop a formula for the area of a triangle given two angles and an included side.

Theorem 10.9 **(ASA Area Formula):** For all triangles ABC,

$$\alpha(\Delta ABC) = \frac{1}{2}a^2\frac{\sin B \cdot \sin C}{\sin(B+C)}.$$

Proof: Our goal is to obtain a formula with one side a and the included angles B and C. From Theorem 10.8,

$$\alpha(\Delta ABC) = \tfrac{1}{2}ac\sin B \cdot \frac{\frac{1}{2}ab\sin C}{\frac{1}{2}bc\sin A}$$
$$= \tfrac{1}{2}a^2\sin B \cdot \tfrac{\sin C}{\sin A}.$$

Since for all x (measured in degrees), $\sin x = \sin(180° - x)$, and since $m\angle A + m\angle B + m\angle C = 180°$, $\sin A = \sin(B + C)$. Substitution of $\sin(B + C)$ for $\sin A$ in (2) yields the desired formula.

We have now shown five formulas for the area of a triangle. Are there others? Of course there are. A formula is possible using any segments or angles whose lengths and measures determine a unique triangle. Problem 16 shows a formula involving the radius of the circumcircle of a triangle. New formulas may be derived by reworking existing formulas. For instance, each line in the proof of Theorem 10.9 yields a different formula for the area of a triangle. The existence of so many ways of calculating the area of a triangle is useful not only because it enables calculation from a variety of given information, but also because the areas of other figures are often calculated by adding areas of triangles.

10.1.2 Problems

1. In the proof of the lemma in this section, demonstrate that $BE = s - AC$.

2. Modify Hero's Formula to yield a formula for the area of a triangle in terms of its sides a, b, and c, and its perimeter p.

3. Refer to the synthetic proof of Hero's Formula.

a. To arrive at step (3), the statement is made that because $LBGC$ is inscribed in a semicircle, its opposite angles are supplementary. Why is this true?

b. Why is $\Delta GKE \sim \Delta LKB$?

c. Where is the geometric mean in step (7)?

4. a. Use Hero's Formula to find the area of a 3-4-5 right triangle.

b. Find the area of the triangle with sides 13, 14, and 15.

c. Use your answer to part **b** to determine the lengths of the three altitudes of the triangle with sides 13, 14, and 15.

d. Your answers to parts **a** and **b** should be integers. Find another triangle whose sides are consecutive integers and whose area is an integer.

5. Use Hero's Formula to prove that if T is a similarity transformation with magnitude k, then, for any triangle ABC, $\alpha(T(\Delta ABC)) = k^2 \cdot \alpha(\Delta ABC)$.

6. Identify all the cyclic quadrilaterals shown in Figure 13 and explain why they are cyclic.

7. Brahmagupta's Formula. The following area formula for cyclic quadrilaterals was known to the Indian mathematician Brahmagupta (c. 598–c. 665) and is named after him. Let Q be a cyclic quadrilateral. Then

$$\alpha(Q) = \sqrt{(s-a)(s-b)(s-c)(s-d)},$$

where the semiperimeter $s = \frac{a+b+c+d}{2}$.

a. Use a geometric construction program to construct a dynamic confirmation of Brahmagupta's Formula for any circle C and any convex quadrilateral Q inscribed in C.

b. Prove that Brahmagupta's Formula does not hold for all quadrilaterals.

c. Explain why Hero's Formula is a special case of Brahmagupta's Formula.[3]

8. If a circle can be inscribed in a cyclic quadrilateral (i.e., all sides of the cyclic quadrilateral are tangent to the circle), then the quadrilateral is called **cyclic-inscribable**. The area of a cyclic-inscribable quadrilateral with sides w, x, y, and z is $\sqrt{wxyz}$. Find a cyclic-inscribable cyclic quadrilateral and show that the formula works for it.

9. Deduce the Law of Sines (Theorem 9.6) using Theorem 10.8.

10. Suppose the area and length of one side of a triangle are fixed. Determine the conditions under which the sum of the lengths of the other two sides is smallest.

11. Suppose the length of one side and the sum of the lengths of the other two sides of a triangle are known. Determine the conditions under which the area is maximized.

12. Prove: Of all triangles with a given perimeter, the equilateral triangle has the greatest area.

13. Use the result of Problem 11 to prove: Of all polygons with a given perimeter, the one with greatest area is equilateral.

14. Find the area of a parallelogram with adjacent sides a and b and *non*-included angle ϕ.

15. a. Find the area of the triangle with consecutive vertices $(0, 0)$, (a, b), and (c, d).

b. Three parallelograms have three vertices (not necessarily consecutive) at (a, b), $(0, 0)$, and (c, d). Find the fourth vertex and area of each parallelogram.

16. Let R be the radius of the circumcircle of ΔABC. Find and write a proof that $\alpha(\Delta ABC) = \frac{abc}{4R}$. (See, for example, Nathan Altschiller-Court, *College Geometry: An Introduction to the Modern Geometry of the Triangle and the Circle*, 2nd edition. New York: Barnes and Noble, 1952.)

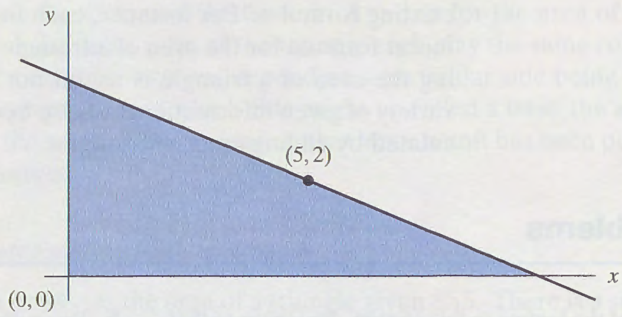
[3] A proof of Brahmagupta's Formula can be found in Howard Eves, *An Introduction to the History of Mathematics with Cultural Connections*, sixth edition (New York: Saunders College Publishing, 1983).

10.1.3 Extended analysis: the line through a given point minimizing area

We devote this section to the analysis of the following problem, based on a problem found in a high school precalculus text.

> Of all lines through the point $(5, 2)$, find the line that cuts off the triangle of smallest area in the first quadrant (Figure 16).

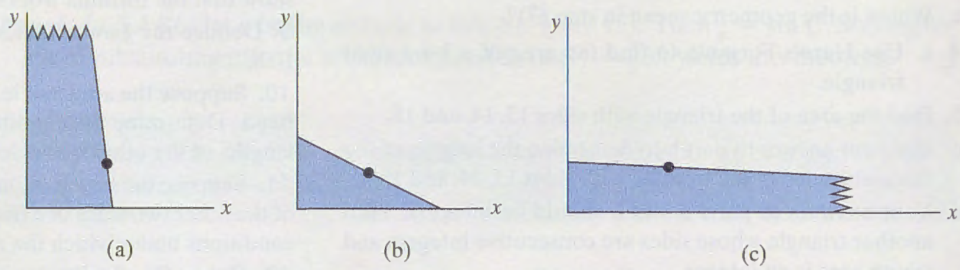
Figure 16



Doing the analysis yourself

You can get a rough feeling about this problem situation by sketching some different lines through the point $(5, 2)$. Notice that a very steep line (Figure 17a) will cut off a large area. Similarly, a very flat line (Figure 17c) will also cut off a large area. (See Problem 1.) It makes sense that there is a line somewhere in the middle range of steepness that cuts off the smallest area.

Figure 17



Before reading beyond this page, do as much as you can of the following.

a. Solve this problem using any method, and justify your solution.

b. Generalize your solution so that, starting with any point in the first quadrant, you could quickly find the line through that point cutting off minimum area.

c. Generalize your solution so that it would apply to a situation where the axes are not perpendicular.

d. Give an alternate approach to the problem that is fully geometrical. (Here a geometrical approach is an approach that does not rely on coordinates to identify the point.)

e. Suggest ways your result could be generalized even further, perhaps in the form of conjectures that you cannot yet see how to prove or disprove.

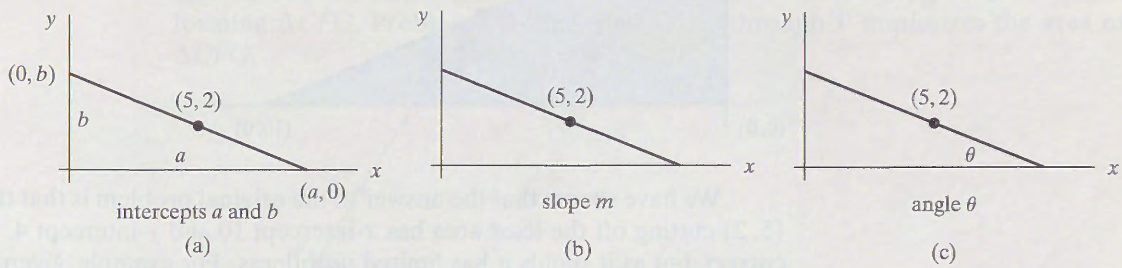
In typical school work, this sort of problem ends at (a). Yet there is a surprising mathematical richness in this situation that remains hidden if we limit ourselves to (a). The purpose of this section is to analyze this problem from an advanced standpoint, generalizing it and extending it along the lines of (a) to (e). Reading the rest of this section will make far more sense if you have at least tried (a)–(e).

The generalizations necessary to do parts (b) and (c) are aided by *parameterization*. We parameterize all the lines through the point (5, 2) by choosing one of the parameters of a line as a variable. The interaction of variables, constants, and parameters is a primary feature of the analysis. The generalizations are also aided by *finding minimal values*. To prove that a particular line through (5, 2) cuts off the smallest area, we use a simple method. We show geometrically that rotating the line a small amount produces a line that cuts off greater area.

Parameterizing the lines

There are several ways to parameterize the lines through (5, 2). Three ways are suggested by Figures 18a, 18b, and 18c.

Figure 18



In Figure 18a, the lines are parameterized by their intercepts a and b. But notice that a and b are not independent. The fact that the line passes though (5, 2) means that a and b are related by

$$(1) \qquad \frac{2-b}{5} = \frac{2}{5-a}.$$

Question 1: Why is (1) true?

The parameterization of Figure 18b, using slope, is discussed in Problem 2. The parameterization using the angle θ, suggested in Figure 18c, is the subject of Problem 1. We show an analysis using Figure 18a.

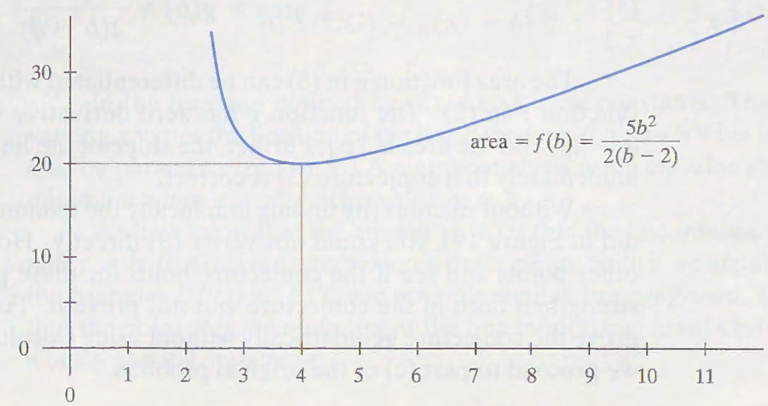
Representing the area

Using the given parameters in Figure 18a, the area of the triangle is $\frac{1}{2}ab$. From the result in (1) we can express a in terms of b. Hence we can represent the area of the triangle in terms of the single parameter b.

$$(2) \qquad \text{area} = f(b) = \frac{5b^2}{2(b-2)}$$

The problem is now reduced to finding the minimum value of the function f. One way to do this is to graph f and look for its minimum value (see Figure 19). From the graph it seems that the minimum of the function occurs when $b = 4$.

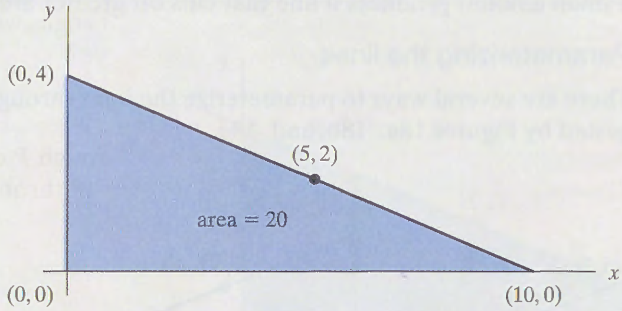
Figure 19



Alternatively, the function (2) can be analyzed using calculus. (See Problem 3.) Doing this verifies the graphical analysis. The minimum occurs precisely when $b = 4$.

We now have an answer to the original problem (a). The line cutting off minimal area has y-intercept $b = 4$. The area it cuts off is $f(4) = 20$. Figure 20 is a diagram to scale.

Figure 20



We have shown that the answer to the original problem is that the line through (5, 2) cutting off the least area has x-intercept 10 and y-intercept 4. This answer is correct, but as it stands it has limited usefulness. For example, given another point, say (3, 7), in the first quadrant and asked to find the line through this point cutting off minimum area, you would have to start the analysis over. At this point, if you have not done part (b) of the original problem, try to do so.

Interpreting the initial result

You may have noticed that there is a simple relationship between the numbers 10 and 4 and the given point (5, 2). We can express this relationship in a general way using the slope.

(3) Of all lines through the point (p, q) in the first quadrant, the slope of the line that cuts off the triangle of smallest area is $-\frac{q}{p}$.

At this point (3) is only a conjecture. However, we can verify this conjecture readily using calculus. The relationships in (1) and (2) based on the specific point (5, 2) can be rewritten in terms of a general point (p, q) as

(4) $$\frac{b-q}{p} = \frac{q}{a-p}$$

and

(5) $$\text{area} = g(b) = \frac{pb^2}{2(b-q)}$$

The area function g in (5) can be differentiated with respect to b as easily as the function f in (2). The function g has zero derivative when $b = 2q$. At this point $a = 2p$, and the area is $2pq$. Further, the slope of the line is $-\frac{q}{p}$. This analysis shows immediately that conjecture (3) is correct.

Without calculus (by finding graphically the minimum point of a function, as we did in Figure 19), you could not verify (3) directly. However, you could try a few other points and see if the conjecture holds for these points. (It will.) This would strengthen faith in the conjecture, but not prove it. Fortunately, there is a way to prove the conjecture geometrically without using calculus. Before showing this way, we proceed to part (c) of the original problem.

Generalization to angles other than right angles

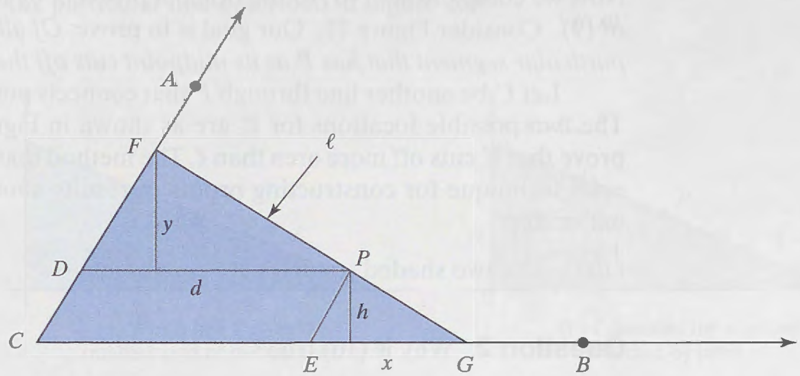
In (c) you are asked to generalize further by looking at a point in any angle, not necessarily a right angle. You should try this problem if you have not done so already.

Here is a way to formulate (c) as a more general problem, which has been posed by Polya.

(6) Given a point P in the interior of an angle, what line through this point forms the triangle with minimum area?

In Figure 21, we have replaced the perpendicular axes with axes meeting to form $\angle ACB$. The line segment $\overline{FG}$ shown through P cuts off a region in this angle, forming ΔCFG. Problem (6) asks what line ℓ through P minimizes the area of ΔCFG.

Figure 21



To proceed we need to identify in some way the location of the point P in the interior of $\angle ACB$. One way to do this is to use the parameters d (= DP) and h (= the distance from P to the side $\overline{CB}$ of the angle), where $\overline{DP} \parallel \overline{CB}$, and $\overline{EP} \parallel \overline{CA}$.

We also need to parameterize the line ℓ. Using x (= EG) and y (the distance from F to $\overline{DP}$) as parameters leads to a very simple analysis. The triangles DFP and EPG are similar. (Why?) Therefore, $\frac{y}{h} = \frac{d}{x}$, so the parameters x and y are not independent.

(7) $$y = \frac{dh}{x}$$

$\alpha(\Delta FCG) = dh + \frac{1}{2}dy + \frac{1}{2}hx$. Using (7), we can represent the area as a function of the parameter x.

(8) $$\alpha(\Delta FCG) = g(x) = h\left(d + \frac{1}{2}\frac{d^2}{x} + \frac{1}{2}x\right).$$

In the function defined by (8), d and h are constants. The variable is x, and varying x varies the position of the line ℓ through P. (This is what it means for the line ℓ to be parameterized by x.) A simple analysis using calculus shows that (8) has a minimum when $x = d$. At this value, $y = h$.

We now know that the answer to (6) is that the line minimizing area has $x = d$ and $y = h$. But this is a cumbersome description. Still, if we are alert we can see that the triangles DFP and EPG are not only similar, but *congruent*. (Why?) This means that the point P is *the midpoint* of the line minimizing area! Our analysis has led to a fairly general statement.

(9) Given a point P inside $\angle ACB$, the line through P that forms the triangle FCG of minimal area, with F on $\overrightarrow{CA}$ and G on $\overrightarrow{CB}$, is the line such that P is the *midpoint* of $\overline{FG}$.

The "midpoint" condition of (9) represents a significant result that, as you will see, is even more general than its formulation here.

Notice that as we generalized this problem, we moved from a result (3), which is described in terms of numbers (coordinates and slopes), to a result (9), which is described in purely geometrical terms. In other words, although the initial formulation and proofs used analytic concepts such as functions and graphs (and calculus), what we discovered is a result about geometry.

Seeing this as purely a geometry problem

Now we complete step (d) of the original problem by looking for a geometric proof of (9). Consider Figure 21. Our goal is to prove: *Of all line segments through P, the particular segment that has P as its midpoint cuts off the minimal area.*

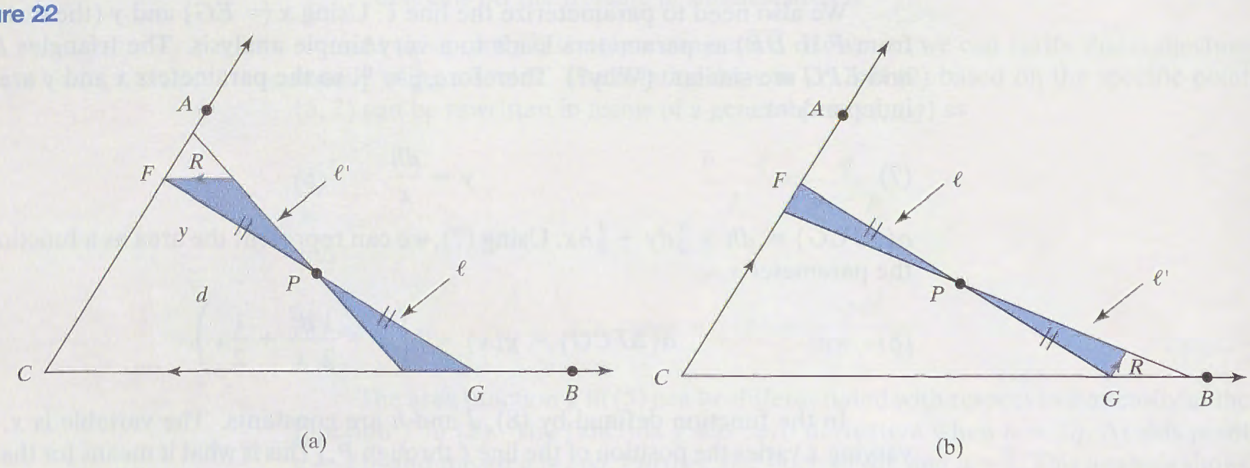
Let ℓ' be another line through P that connects points on two sides of the angle. The two possible locations for ℓ' are as shown in Figures 22a and 22b. We try to prove that ℓ' cuts off more area than ℓ. The method that we use is a general and powerful technique for constructing proofs of results about optimization. Here is the major step:

(10) The two shaded triangles are congruent.

Question 2: Why is (10) true?

From (10), we see that the area enclosed by line ℓ' is greater than the area enclosed by line ℓ. In fact, it is greater by exactly the area of region R.

Figure 22



Summary of what we have done

In summary, we have constructed an extended analysis of the original high school problem following the steps of (a) to (d). The analysis resulted in significant generalizations. In particular, we generalized beyond particular numbers (the point (5, 2)), to a general point (p, q). We also generalized beyond particular shapes (a right angle) to any angle. The analysis progressed away from the original mathematical subject area of the problem (coordinate systems, graphs, and functions) to one more natural for the problem at hand. Finally, the more general analysis is no more difficult to carry out than the original very specific one.

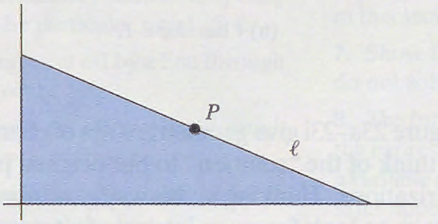
The last feature is especially noteworthy. Although (9) is a very general result, the geometric proofs illustrated in Figures 22a and 22b are easier than the earlier analysis in terms of graphs and functions.

What about step (e) of the original problem? As an example, based on (9) we might conjecture that the line ℓ through a point P inside a parabola that cuts off the smallest area is the line bisected by P. (See Project 1 at the end of this chapter.)

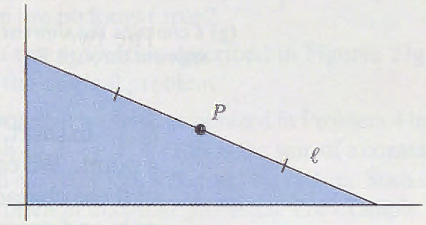
What to look for in advance in general solutions

There is an important further perspective on what we have done in our analysis. In briefest terms, we have shown that the question asked in Figure 23a is answered by the particular line described in Figure 23b.

Figures 23a, b



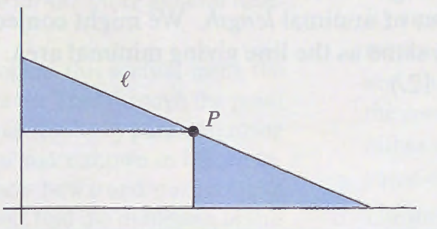
(a) Which line ℓ gives the smallest area of the triangle?



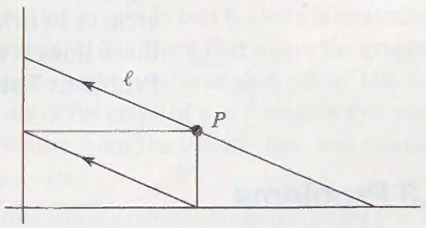
(b) ℓ contains the segment bisected by point P.

Let us call the line described in Figure 23b a "special" line. It is special in the sense that it has a simple geometric characterization in terms of the given point P. In Figures 23c through 23i, we indicate seven other conditions on lines through P.

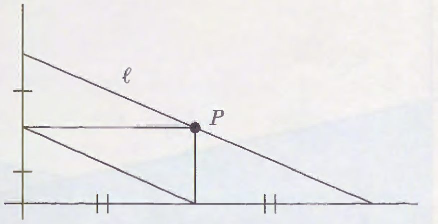
Figures 23c–f



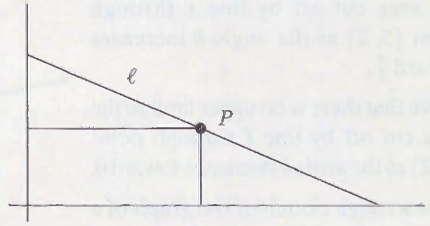
(c) ℓ gives triangles of *equal area* (shaded).



(d) ℓ is *parallel* to the diagonal of the rectangle.

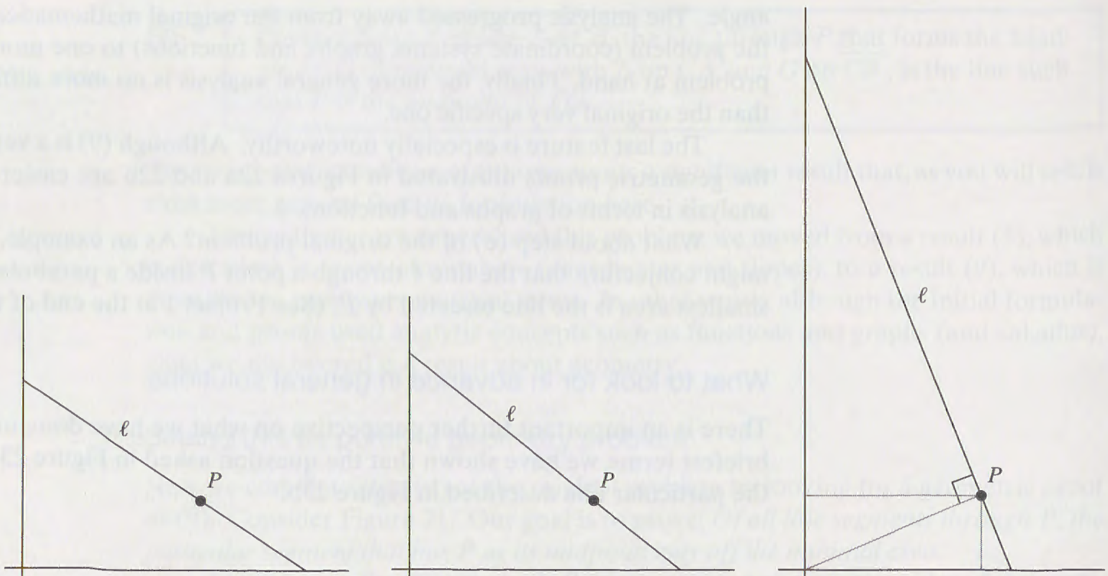


(e) ℓ has intercepts that are *double* the sides of the rectangle.



(f) ℓ creates the triangle *with the smallest perimeter.*

Figures 23g–i



(g) ℓ contains the *shortest segment* through P.

(h) ℓ has *slope -1*.

(i) $\ell \perp \overline{OP}$

In short, Figure 23a–23i give geometric ways of characterizing special lines through a point. We can think of the "solution" to our original problem as being, possibly, one of these characterizations. Thinking in this way gets us away from focusing on numbers only and prompts us to look for more intrinsic features of our solutions.

Our analysis showed that Figures 23a and 23b are equivalent. This is the substance of result (9). What about the other "special" lines of Figure 23? It turns out that the five characterizations in Figures 23a to 23e are all equivalent: If a line has one of these properties, it has them all. (See Problem 7.)

In some cases, when people are asked to use intuition to guess what line gives minimum area, they come up with the properties described in Figures 23f through 23i. These are reasonable guesses, but further analysis shows them to be not correct. Figures 23f and 23g describe two other minimal properties, this time not minimal *area*, as in (a), but of minimal *length*. We might conjecture that one or the other of these lines is the same as the line giving minimal area. But this is not the case. (See Problems 7 and 12.)

10.1.3 Problems

1. Use Figure 24.

a. Prove that there is no upper limit to the area cut off by line ℓ through point $(5, 2)$ as the angle θ increases toward $\frac{\pi}{2}$.

b. Prove that there is no upper limit to the area cut off by line ℓ through point $(5, 2)$ as the angle θ decreases toward 0.

c. Give a rough sketch of the graph of a function that represents the area of ΔOAB as a function of θ.

Figure 24

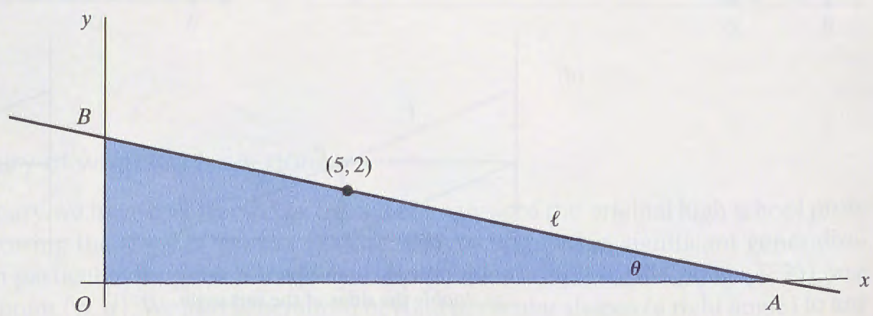
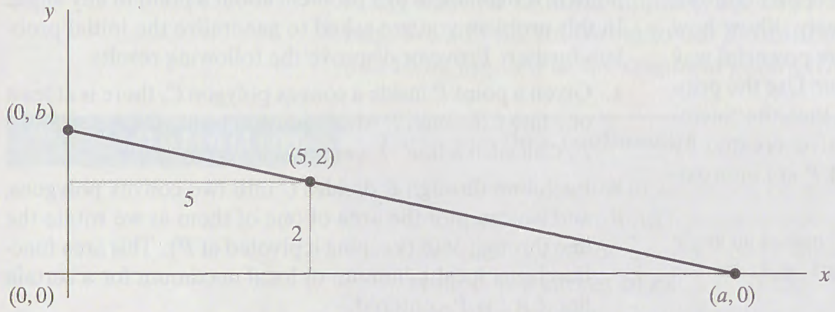


Figure 25



2. A nonvertical line in $\mathbf{R}^2$ is associated with three basic parameters: its slope m, its y-intercept b, and its x-intercept a. (See Figure 25.)

a. Show that any *two* of these parameters suffice to specify any nonvertical line.

b. Show that any *one* of these parameters suffices to specify any nonvertical line through the particular point $(5, 2)$.

c. Represent the area of the triangle cut off by a line through $(5, 2)$ in terms of the *slope m* of the line.

d. Represent the area of the triangle cut off by a line through $(5, 2)$ in terms of the *y-intercept b* of the line.

e. Represent the area of the triangle cut off by a line through $(5, 2)$ in terms of the *x-intercept a* of the line.

3. a. Give the details of the calculus proof that the function $f: b \to \frac{5b^2}{2(b-2)}$ of Figure 19 obtains its minimum at $b = 4$.

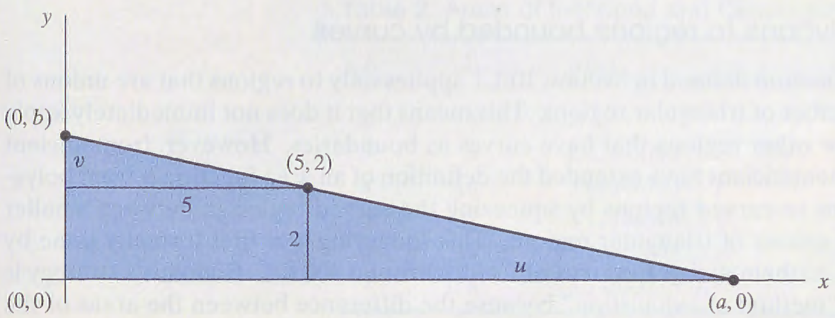
b. Give the details of the calculus proof that the more general function $f: b \to \frac{pb^2}{2(b-q)}$ obtains its minimum at $b = 2q$.

c. Use polynomial division to find an equation for the oblique asymptote of the graph of the function $f: b \to \frac{5b^2}{2(b-2)}$.

d. Answer the question of part **c** for the more general function of part **b**.

4. We solved the problem that opens this section using the intercepts a and b to parameterize the lines through the point $(5, 2)$. A somewhat simpler analysis results by parameterizing the lines in terms of the lengths u and v shown in Figure 26. Carry out this analysis. That is, show how u and v are related, express the area in terms of v alone, find the minimum of this function, and answer the original problem of this section.

Figure 26



5. We have given a simple characterization of the segment through a point P that cuts off minimal area in the first quadrant: It is the segment bisected by P. Show that the other characterizations of the minimal area in Figures 23a to 23e are equivalent.

6. Suppose the point (p, q) in relationship (3) of this section is in the 2nd quadrant. Which, if any, of relationships (3) to (5) in this section are no longer true?

7. Show that the properties described in Figures 23g to 23i do not solve the original problem.

8. The function that we have minimized in Problem 4 involves the expression $10 + \frac{5}{2}v + \frac{10}{v}$. This is the sum of a constant and a variable and its reciprocal, with constant factors. Such expressions appear often in max-min problems. For example, such a function appears in (8). This problem asks that you show that there is a general answer to all such problems:

a. Consider a function $f(x) = Ax + \frac{B}{x}$, where $A > 0$ and $B > 0$ are constants. Find the minimum value of such a function, and find where it achieves this minimum. (Use calculus.)

b. Show that the minimum value of $Ax + \frac{B}{x}$ occurs at the value of x where the graph of $y = Ax$ intersects the graph of $y = \frac{B}{x}$.

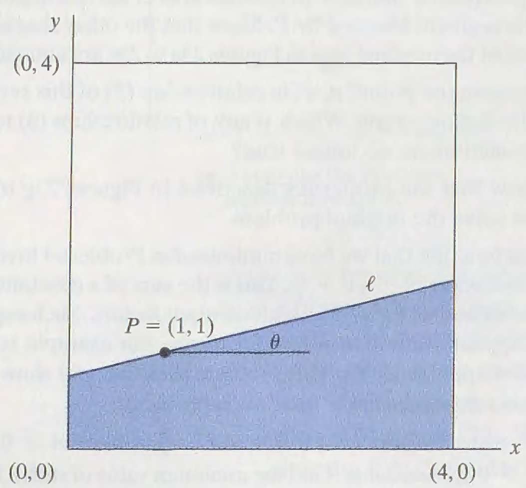
c. By analyzing the graphs in part **b**, give a noncalculus argument for part **a**. (*Hint:* Show that where the graphs intersect, the slopes are opposites of each other. Use that and the concavity of the graph of $y = \frac{B}{x}$ to show that moving in either direction from the intersection will increase the value of the sum.)

d. Use the Arithmetic-Geometric Mean Inequality ($\frac{a+b}{2} \geq \sqrt{ab}$ for all real a and b) to do parts **a** and **b**.

*9. In this section, we *verified* that the line through point P cutting off minimum area has the midpoint property. But we had already conjectured the midpoint property. Show how the method we used can also be used in a more powerful way to actually *derive* the midpoint property. (*Hint*: Use the principle that if a line ℓ does cut off minimal area, then the "increments" of area, one positive and one negative, created by rotating this line a small amount about point P are approximately equal.)

10. A line ℓ passing through point $P = (1, 1)$ makes an angle θ with the horizontal. (See Figure 27.) Let $A(\theta)$ be the shaded area below ℓ in the square.

Figure 27



a. Sketch a graph of $y = A(\theta)$ as θ ranges from 0 to 2π.
b. Find a formula for $A(\theta)$ and graph this function. (There may need to be several formulas, depending on which sides of the square the line ℓ intersects.)

*11. In this section, we extended the initial problem about a point in a right angle to a problem about a point in any angle. In this problem you are asked to generalize the initial problem further. Prove or disprove the following results.

a. Given a point P inside a convex polygon C, there is at least one line ℓ through P whose portion inside C is bisected by P. Call such a line "P-centered".
b. Each line through P divides C into two convex polygons, and we can plot the area of one of them as we rotate the line through 360° (keeping it pivoted at P). This area function has a local minimum or local maximum for a certain line ℓ if ℓ is P-centered.

12. Given a point (m, n) in the first quadrant,

a. Use calculus to show that the *shortest* line segment in the quadrant through (m, n) has slope $\left(\frac{n}{m}\right)^3$. (*Hint*: Parameterize the lines by the angle they make with the x-axis.)
b. Relate this problem to the problem of finding the shortest ladder that will touch the wall, the floor, and a box of height n and width m placed against the wall.
c. Relate this problem to the problem of finding the *longest* ladder that will fit around a corner where two hallways of widths m and n meet.
*d. Use a geometric argument to derive the property stated in part **a**.

13. Let P be a fixed point in the interior of a parabola. Let ℓ be a line containing P and intersecting the parabola at points A and B. What position of ℓ minimizes the area of the region bounded by ℓ and the parabola? (*Hint*: Because all parabolas are similar, answering the question for one parabola essentially answers the question for all parabolas.)

*14. Investigate the situation of the line ℓ through a fixed point P in the first quadrant outlining, with the axes, the triangle of minimal perimeter. (See Figure 23f.)

ANSWERS TO QUESTIONS

1. Each side of the equation is the slope determined by $(5, 2)$ and one of the intercepts.

2. In each case the midpoint P determines a pair of congruent sides, a pair of angles are vertical angles, and another pair of angles are alternate interior angles formed by two parallel lines. So the triangles are congruent by ASA congruence.

10.1.4 From polygons to regions bounded by curves

The area function defined in Section 10.1.1 applies only to regions that are unions of a finite number of triangular regions. This means that it does not immediately apply to circles or other regions that have curves as boundaries. However, from ancient times, mathematicians have extended the definition of an area function α from polygonal regions to curved regions by squeezing the curved region in between smaller and larger unions of triangular regions. This squeezing was first formally done by the Greek mathematician Eudoxus of Cnidus around 400 B.C. Eudoxus's strategy is called the "method of exhaustion" because the difference between the areas of the smaller and larger unions is exhausted in the sense that it approaches zero. It is akin to the way the rational numbers are extended to the real numbers via nested intervals. We add the following to our definition of area to enable the method of exhaustion to be applied to the common geometric figures with curved boundaries.

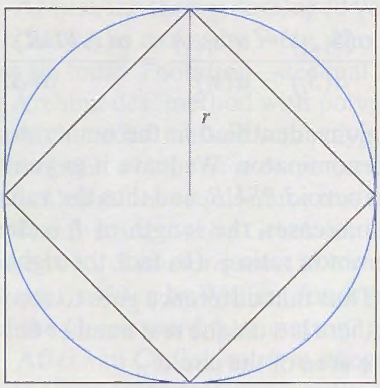
Definition **(of area function, continued):**

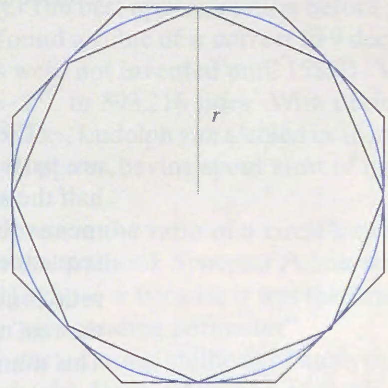
4. Let C be a region in E^2. Let s_i be a sequence of unions of triangular regions in E^2 with no interior points in common, such that each s_i is a subset of C. Let S_i be a sequence of unions of triangular regions in E^2 such that C is a subset of each S_i. If the least upper bound of the $\alpha(s_i)$ equals the greatest lower bound of the $\alpha(S_i)$, then this bound is $\alpha(C)$.

Approaching the area of a circle from above and below

We use property (4) to obtain the area of a circle. Figure 28a shows inscribed and circumscribed squares in a circle. The area of the circle is between the areas of the squares. If the radius of the circle is r, then Figure 28a shows that the area must be between $2r^2$ and $4r^2$. Figure 28b uses inscribed and circumscribed octagons, whose areas are $2\sqrt{2}r^2$ and $8(\sqrt{2}-1)r^2$, that is, between approximately $2.8284r^2$ and $3.3137r^2$.

Figure 28



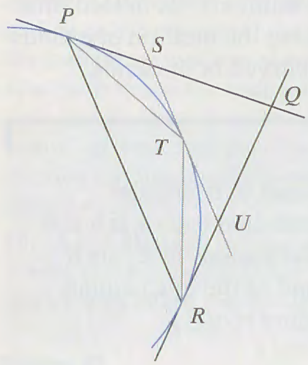


The calculations in Table 2 confirm that these areas continue to get closer to each other as the number of sides of the polygons is doubled again and again. In Table 2, s_i is a regular inscribed polygon with 2^i sides (a 2^i-gon) in a circle of radius 1, where $i \geq 2$, and S_i is the corresponding circumscribed polygon. Thus each polygon in both sequences has twice the number of sides of the preceding term of that sequence. Because each inscribed polygon can be thought of as connecting the points of tangency of the circumscribed polygon, for all i, $\alpha(s_i) \leq \alpha(S_i)$.

Table 2 Areas of Inscribed and Circumscribed 2^i-gons

i	2^i	$\alpha(s_i)$	$\alpha(S_i)$	$\alpha(S_i) - \alpha(s_i)$
2	4	$2r^2$	$4r^2$	$2r^2$
3	8	$2.82842711\ldots r^2$	$3.31370849\ldots r^2$	$.4852813\ldots r^2$
4	16	$3.06146745\ldots r^2$	$3.18259787\ldots r^2$	$.1211304\ldots r^2$
5	32	$3.12144515\ldots r^2$	$3.15172490\ldots r^2$	$.0302797\ldots r^2$
6	64	$3.13654849\ldots r^2$	$3.14411838\ldots r^2$	$.0075698\ldots r^2$
7	128	$3.14033115\ldots r^2$	$3.14222363\ldots r^2$	$.0018924\ldots r^2$

Figure 29



In Figure 29 we show parts of regular inscribed and circumscribed 2^i-gons and 2^{i+1}-gons. $\overline{PR}$ is a side of a regular inscribed 2^i-gon and $\overline{PQ}$ and $\overline{QR}$ are halves of sides of a regular circumscribed 2^i-gon. $\overline{PT}$ and $\overline{TR}$ are sides of a regular inscribed 2^{i+1}-gon, while $\overline{PS}$, $\overline{UR}$, and $\overline{SU}$ are two halves and a full side of a regular circumscribed 2^{i+1}-gon. The figure guides us to see that, as the number of sides is doubled, $\alpha(s_{i+1}) > \alpha(s_i)$ because $\alpha(\Delta PTR)$ is added to the area, while $\alpha(S_{i+1}) < \alpha(S_i)$ because $\alpha(\Delta SQU)$ is taken away. Thus we have, for each i,

$$\alpha(s_i) < \alpha(s_{i+1}) < \alpha(S_{i+1}) < \alpha(S_i).$$

So the closed intervals $I_i = [\alpha(s_i), \alpha(S_i)]$ form a nested sequence.

Figure 29 also shows us how the lengths of the intervals I_i go to zero as i increases. The length of I_i is $\alpha(S_i) - \alpha(s_i)$, the difference between the areas of the circumscribed and inscribed 2^i-gons. From Figure 29, which shows only $\frac{1}{2^i}$ of the circle,

$$\alpha(S_i) - \alpha(s_i) = 2^i(\alpha(\Delta PQR))$$

and

$$\begin{aligned}\alpha(S_{i+1}) - \alpha(s_{i+1}) &= 2^{i+1}(\alpha(\Delta TUR))\\ &= 2^i(\alpha(\Delta TUR) + \alpha(\Delta TSP)).\end{aligned}$$

Consequently,

$$\frac{\alpha(S_{i+1}) - \alpha(s_{i+1})}{\alpha(S_i) - \alpha(s_i)} = \frac{\alpha(\Delta TUR) + \alpha(\Delta TSP)}{\alpha(\Delta PQR)}.$$

The two regions identified in the numerator of the fraction on the right side are parts of the denominator. We leave it to you to show that their sum is less than half the area of trapezoid $PSUR$, and thus the value of the fraction is less than $\frac{1}{2}$. This means that, as i increases, the length of I_i is decreasing faster than a geometric sequence with common ratio $\frac{1}{2}$. (In fact, the right column of Table 2 shows that the ratio is nearer $\frac{1}{4}$.) Thus that difference goes to zero. By the completeness property of the real numbers, there is a unique real number between all the $\alpha(s_i)$ and all the $\alpha(S_i)$. This number is the area of the circle.

π and πr^2

But how do we know that the area of a circle with radius r is πr^2? This is perhaps the easiest part of the entire argument. We define π in its usual way, as the ratio of the circumference C of the circle to its diameter. That is, $\pi = \frac{C}{d}$. From this, $C = \pi d = 2\pi r$.

Theorem 10.10 The area of a circle with radius r is πr^2.

Proof: When a regular 2^n-gon circumscribes a circle of radius r, the circle is inscribed in the polygon. By Theorem 10.6, the area of this polygon is given by $A = \frac{1}{2}rp$. As n increases, the perimeter p is getting closer and closer to the perimeter of the inscribed 2^n-gon, and in between these perimeters is the circumference of the circle. So the area of the circle is between two areas, each of which is approaching the value $\frac{1}{2}rC$, that is, $\frac{1}{2}r \cdot 2\pi r$, which is πr^2. ∎

Thus the area formula for a circle is an extension of the area formula for polygons that can be circumscribed about a circle. The number π enters the formula because π is defined in terms of the "perimeter" of the circle, that is, in terms of the circumference.

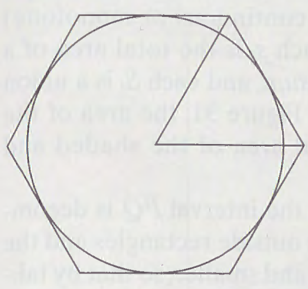


Figure 30

The connection between circumference and area, as mentioned in the proof of Theorem 10.10, was first made by Archimedes using the method of exhaustion. In doing so, Archimedes also was able to find a rather good approximation to π. Here is what he did. Instead of starting with inscribed and circumscribed squares as we did, he began with the triangle at the center of the circle that is $\frac{1}{6}$ of a circumscribed regular hexagon (see Figure 30). He successively bisected the central angle, compared ratios, took away parts of irrational square roots, and arrived at the conclusion that the circumference of a circle is less than $3\frac{1}{7}$ times the diameter.

Then Archimedes considered inscribed regular polygons of 6, 12, 24, 48, and 96 sides. He found the perimeter of each polygon and concluded that the circumference of a circle is more than $3 + \dfrac{1}{7 + \frac{1}{10}}$ times the diameter. In this way, Archimedes provided the first good numerical approximation to π: $3\frac{10}{71} < \pi < 3\frac{1}{7}$.

A brief history of π

Since the time of Archimedes, the calculation of better and better numerical approximations to π has occupied the attention of mathematicians throughout the world. At least two entire books are devoted to this history— Beckmann's *A History of Pi*, and Berggren and the Borweins *Pi: A Source Book*. We give only the briefest of histories of this calculation.

At first, the theory developed very slowly. The best approximation before 1600 that is known to us was by Viéte, who in 1576 found a value of π correct to 9 decimal places (in today's notation—decimal fractions were not invented until 1585!). Viéte used Archimedes' method with polygons of $6 \cdot 2^{16}$, or 393,216 sides. With decimals, using Archimedes' method with polygons of 2^{62} sides, Ludolph van Ceulen in 1610 was able to obtain a value of π correct to 35 decimal places, having spent most of his life on the calculations required for this task.

The first use of the Greek letter π to represent the ratio of a circle's circumference to its diameter seems to have been in the textbook *Synopsis Palmariorium Mathesios*, written by William Jones in 1706. He chose π because it was the first letter of the Greek word "perimetrog", meaning "surrounding perimeter".

After van Ceulen, most mathematicians started using methods of analysis and infinite series to approximate π. By 1844, Zacharias Dase, a German of prodigious calculating ability, computed π to 200 decimal places. And in 1873, with 15 years of work, William Shanks computed π to 707 places, of which the first 527 were correct. This was the most accurate calculation before the days of machine calculation.

Shanks's error was found in 1948 by mathematicians using a desk calculator. The next year, one of the very earliest computers, ENIAC, calculated π to 2,037 decimal places. By 1967, π had been calculated to 500,000 places. Two million places were calculated by Kazunori Miyoshi and Kazuhiko Nakayama in 1981 using the trigonometric identity

$$\pi = 32 \tan^{-1}\left(\frac{1}{10}\right) - 4 \tan^{-1}\left(\frac{1}{239}\right) - 16 \tan^{-1}\left(\frac{1}{515}\right).$$

Within five years, the billion-place standard was reached by David Bailey and Jonathan and Peter Borwein using a formula discovered by Ramanujan in 1910. By the year 2000, more than 51 billion decimal places of π were known.

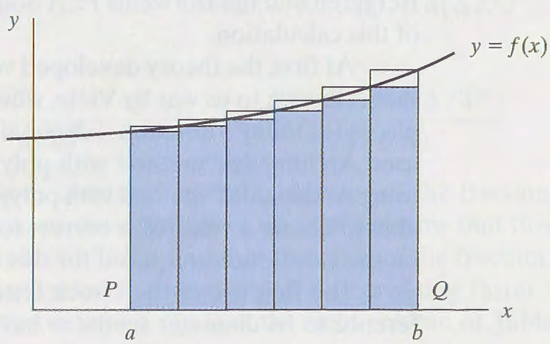
The Riemann integral

In the 17th century, mathematicians refined the method of exhaustion in their development of calculus. If a curve can be described by a formula, calculus often enables the lengthy calculations of the method of exhaustion to be replaced by the relatively simple calculation of definite integrals.

In calculus, the area property (4) shows one way to obtain the area between the graph of a sufficiently well-behaved function (e.g., continuous or monotone) $y = f(x)$, the x-axis, and the lines $x = a$ and $x = b$. Each s_i is the total area of a union of rectangles whose total area is a *lower Riemann sum*, and each S_i is a union of rectangles whose area is an *upper Riemann sum*. In Figure 31, the area of the shaded region is a lower Riemann sum while the total area of the shaded and unshaded rectangles is an upper Riemann sum.

If we increase the number of partitions into which the interval $\overline{PQ}$ is decomposed, the difference between the sum of the areas of the outside rectangles and the sum of the areas of the inside rectangles becomes smaller and smaller, so that by taking a sufficiently large number of partitions, we can make this difference as small as we please. Since the area under the curve lies between these two sums, it is the limit toward which the sum of the outside or inside rectangles tends as the number of partitions is definitely increased, and the determination of this limit is accomplished by what we know from calculus as integration. In other words, the common bound on these sums for a monotone increasing function $y = f(x)$ is the desired area, the value of the definite integral $\int_a^b f(x)\,dx$.

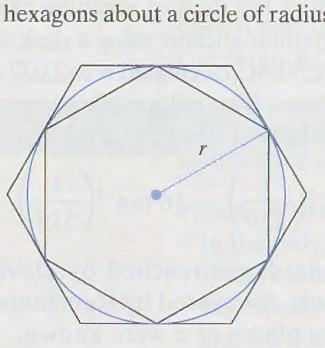
Figure 31



10.1.4 Problems

1. a. Find the exact areas of regular circumscribed and inscribed hexagons about a circle of radius r (Figure 32).

Figure 32



b. If a side of a regular n-gon inscribed in a circle of radius 1 has length x, determine the length of a side of a regular $2n$-gon inscribed in the same circle in terms of x.

*c. Use parts **a** and **b** and a calculator or computer to calculate the perimeter of regular polygons of 12, 24, 48, 96, and 192 sides inscribed in a circle of diameter 1. (Because of the unwieldy number system used at the time, Archimedes could go only so far as 96 sides.)

2. **The area of a sector.** A **sector** is the region bounded by two radii $\overline{OA}$ and $\overline{OB}$ and arc $\widehat{AB}$ of circle O (Figure 33). Assume that the area of a sector is proportional to the measure of the arc. Use this assumption to deduce the formula $A = \frac{r^2\theta}{2}$ for the area of a sector bounded by an arc of radian measure θ in a circle of radius r.

Figure 33

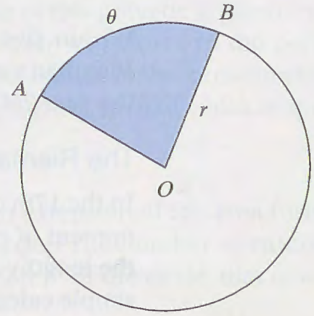
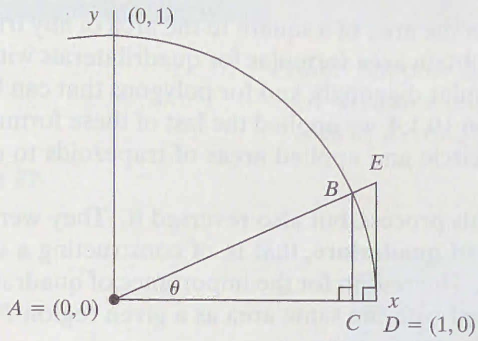


Figure 34



3. Figure 34 shows a part of the unit circle in the first quadrant.

a. Use the result of Problem 2 to prove that $\sin\theta < \theta < \tan\theta$, for θ in radians.

b. Explain why $\lim\limits_{\theta\to 0}\frac{\sin\theta}{\theta} = \lim\limits_{\theta\to 0}\frac{\tan\theta}{\theta} = 1$.

4. Find formulas for the area and perimeter of a Reuleaux triangle of width w. (See Problem 6 of Section 10.1.1.)

5. The area of an ellipse. Let T be the transformation in $\mathbf{R}^2$ defined by $T(x, y) = (ax, by)$.

a. Show that if $JKLM$ is any square in $\mathbf{R}^2$, and $T(JKLM) = J'K'L'M'$, then $\alpha(J'K'L'M') = ab\alpha(JKLM)$.

b. Show that the image of the unit circle under T is the ellipse with equation $\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1$.

c. Use parts **a** and **b** to derive a formula for the area of any ellipse in terms of its semimajor axis a and its semiminor axis b.

6. Area under a parabola. Consider the region R bounded by the parabola $y = x^2$, the x-axis, and the line $x = 10$.

a. Calculate $L = \sum\limits_{i=0}^{9} i^2$ and $U = \sum\limits_{i=1}^{10} i^2$. Explain why $L < \alpha(R) < U$.

b. Find sums that provide a smaller interval containing $\alpha(R)$.

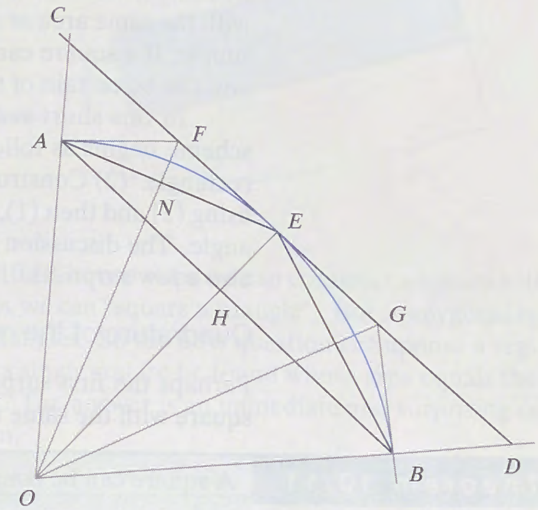
c. Use calculus to determine $\alpha(R)$ exactly.

7. Estimation of π using perimeters. In this section, π was estimated using the areas of inscribed and circumscribed regular polygons. The following theorem enables the estimation of π by calculating successive harmonic and geometric means of the perimeters of these polygons.

> **Theorem:** Let p and P be the perimeters of inscribed and circumscribed regular n-gons in a circle. Let p' and P' be perimeters of inscribed and circumscribed regular $2n$-gons in the same circle. Then
>
> (1) $$P' = \frac{2pP}{p+P} = H(p, P)$$
>
> (2) $$p' = \sqrt{pP'} = G(p, P').$$

That is, suppose the perimeters of the inscribed and circumscribed n-gons of a circle are known. Then the perimeter of the circumscribed $2n$-gon is the harmonic mean of these perimeters, and the perimeter of the inscribed $2n$-gon is the geometric mean of the perimeter of the circumscribed $2n$-gon and the inscribed n-gon. These two recursion relations enable π to be estimated as close as desired (given the computational wherewithal).

Figure 35



In Figure 35, $\overline{AB}$ is a side of a regular n-gon inscribed in circle O and $\overline{CD}$ is a side of the circumscribed regular n-gon to circle O, tangent to the arc $\widehat{AB}$ at its midpoint E. Then $\overline{AE}$ and $\overline{BE}$ are sides of a regular $2n$-gon inscribed in circle O. Tangents at A and B intersect $\overline{CD}$ at F and G, respectively. Then $\overline{FG}$ is a side of a circumscribed regular $2n$-gon in circle O. From these constructions, by definition of *perimeter*, $p = n \cdot AB$ and $P = n \cdot CD$. Also, $p' = 2n \cdot AE$ and $P' = 2n \cdot FG$.

a. Calculate p and P if $n = 2$ and the radius of the circle is 0.5.

b. Use the results of part **a** to calculate successive perimeters of regular inscribed and circumscribed polygons with 2^i-sides until π is estimated correct to 6 decimal places.

c. Prove both parts of the theorem by justifying each of these conclusions.

i. $\frac{P}{p} = \frac{CF}{FE}$ ii. $\frac{P+p}{2p} = \frac{CE}{FG}$ iii. $\frac{P}{P'} = \frac{CE}{FG}$

iv. $\frac{P+p}{2p} = \frac{P}{P'}$, from which (1) follows by solving for P'

v. $\frac{p}{p'} = \frac{AH}{AE}$ vi. $\frac{p'}{P'} = \frac{EN}{EF}$

vii. $\Delta ENF \sim \Delta AHE$, from which $\frac{AH}{AE} = \frac{EN}{EF}$

viii. $\frac{p}{p'} = \frac{p'}{P'}$, from which (2) follows by solving for p'.

10.1.5 The problem of quadrature

In Section 10.1.1, we proceeded from the area of a square to the area of any triangle, and then used areas of triangles to obtain area formulas for quadrilaterals with parallel sides (trapezoids) or perpendicular diagonals, and for polygons that can be circumscribed about a circle. In Section 10.1.4, we applied the last of these formulas to obtain a formula for the area of a circle and applied areas of trapezoids to obtain formulas for areas under curves.

The ancient Greeks utilized this process, but also reversed it. They were particularly interested in the problem of **quadrature**, that is, of constructing a square with the same area as a given figure. The reason for the importance of quadrature is simple: If a square can be constructed with the same area as a given region F, then you can be certain of the area of F in square units.

In this short section, we consider the question of quadrature. Our general scheme begins as follows. (1) Construct a square with the same area as any given rectangle. (2) Construct a rectangle with the same area as any given triangle. Then, using (2) and then (1), we can construct a square with the same area as any given triangle. The discussion provides a review of some of the ideas in earlier sections and also a few surprises.

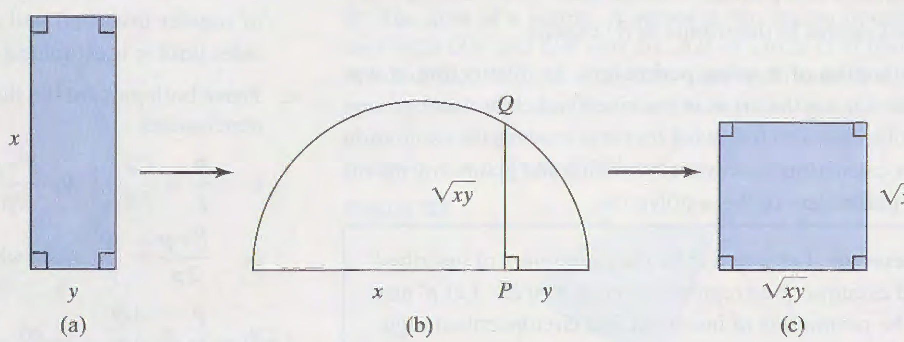
Quadrature of the rectangle

Perhaps the first surprise is that we use ideas from similar triangles to construct a square with the same area as a given rectangle.

Theorem 10.11 A square can be constructed with the same area as a given rectangle.

Proof: Suppose a rectangle has dimensions x and y as in Figure 36a. Then a square with side s and the same area as the quadrilateral has area $s^2 = xy$, so $s = \sqrt{xy}$. Thus s is the geometric mean of x and y. A segment of length s can be constructed from segments of lengths x and y using ideas from Section 8.3.1. Specifically, place segments of lengths x and y on the same line next to each other, with point P in common, as shown in Figure 36b. Construct a circle whose center is the midpoint of the new segment of length $x + y$ and whose radius is $\frac{x+y}{2}$. Then construct the perpendicular at P. The length of the half-chord $\overline{PQ}$ from P to the circle has length $\sqrt{xy}$ and so is the length of a side of the desired square. ∎

Figure 36



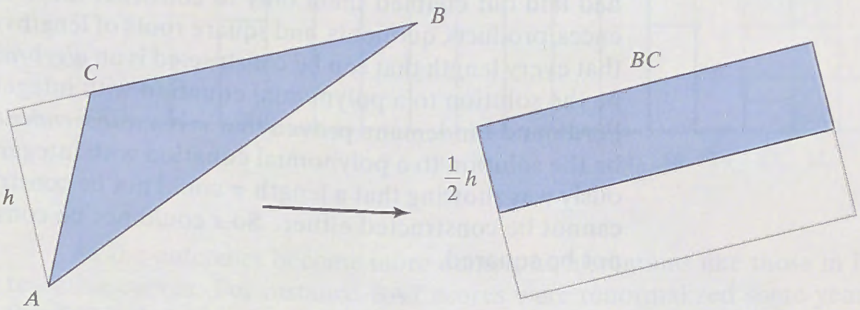
Question: Why does $\overline{PQ}$ have length $\sqrt{xy}$?

We say that Theorem 10.11 allows us to "square a rectangle". Theorem 10.11 is exceedingly important in the theory of quadrature because it is easy to construct a rectangle with the same area as any triangle.

Theorem 10.12 A rectangle can be constructed with the same area as a given triangle.

Proof: Suppose ΔABC is given. Let h be the altitude to side $\overline{BC}$. Then $\alpha(\Delta ABC) = \frac{1}{2}h \cdot BC$, so a rectangle with consecutive sides of lengths $\frac{1}{2}h$ and BC has the same area as ΔABC. Such a rectangle is easy to construct (Figure 37).

Figure 37

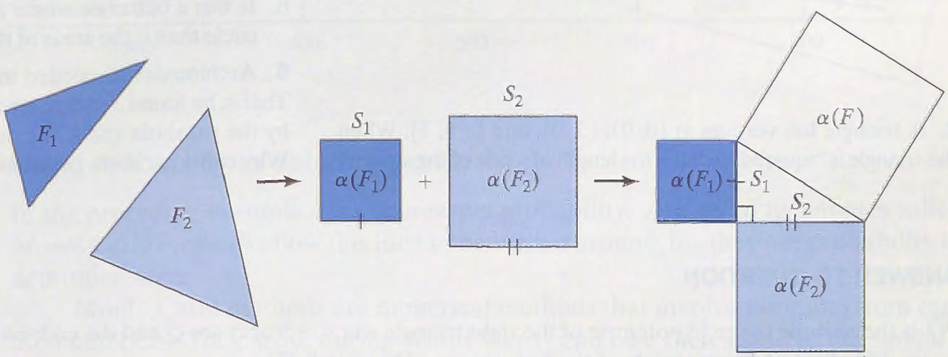


From Theorems 10.11 and 10.12, in two steps we can construct a square with the same area as any triangle. That is, we can "square a triangle". But a polygonal region may be split into two or more triangles. So the next question is: Suppose a region is the union of two triangles. Can a single square be found whose area equals the sum of the areas of the two triangles? The answer is an immediate and surprising consequence of a well-known theorem.

Theorem 10.13 Let F be the union of two triangular regions with no interior points in common. Then a square can be constructed with the same area as F.

Proof: Let the two triangular regions be F_1 and F_2 (Figure 38). By Theorems 10.11 and 10.12, a square region of side s_1 can be constructed with area $\alpha(F_1)$, and a square region of side s_2 can be constructed with area $\alpha(F_2)$.

Figure 38



Now construct a right triangle with legs s_1 and s_2. By the Pythagorean Theorem (!), the square on the hypotenuse has area equal to $\alpha(F_1) + \alpha(F_2)$, so that square has area $\alpha(F)$.

The following corollary can be proved by mathematical induction. It shows that every polygonal region can be "squared". We leave its proof to you.

Corollary: Let F be the union of n triangular regions, no two of which have any interior points in common. Then a square can be constructed with the same area as F.

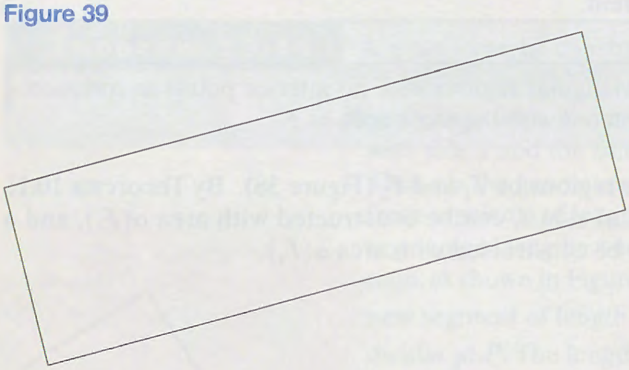
Armed with the knowledge of all the theorems in this section, it was natural for the Greeks to wonder if a square could be constructed with the same area as that of a given circle. But they were unable to solve the problem of "squaring a circle".

Not until the 19th century was the reason for their difficulty established. To square a circle, a side s must be constructed with $s^2 = \pi r^2$, where r is known. That implies that $s = r\sqrt{\pi}$. The rules of ruler-and-compass construction the Greeks had laid out enabled them only to construct finite combinations of sums, differences, products, quotients, and square roots of lengths of given segments. It follows that every length that can be constructed is an *algebraic* number, a number that can be the solution to a polynomial equation with integer coefficients. When in 1882 Ferdinand Lindemann proved that π is a *transcendental* number—one that cannot be the solution to a polynomial equation with integer coefficients—he simultaneously was showing that a length π could not be constructed. This implies that $\sqrt{\pi}$ cannot be constructed either. So s could not be constructed, and so a circle cannot be squared.

10.1.5 Problems

1. Trace the rectangle shown in Figure 39. Using a straightedge and compass, construct a square with the same area.

Figure 39



2. A triangle has vertices at $(0, 0)$, $(5, 0)$, and $(-6, 7)$. When this triangle is "squared", what is the length of a side of the square?

3. A trapezoid has vertices at $(0, 0)$, $(a, 0)$, (b, c), and (d, c). When this trapezoid is "squared", what is the length of a side of the square?

4. Prove the Corollary to Theorem 10.13.

5. A reasonable way to try to square a given circle is as follows. Circumscribe a square about the circle. Inscribe a second square in the circle. Let a third square have a side length equal to the arithmetic mean of the sides of the inscribed and circumscribed squares.

a. Compare the area of the third square to the area of the circle.

b. Is this a better or worse approximation to the area of the circle than if the areas of the first two squares are averaged?

6. Archimedes succeeded in the quadrature of the parabola. That is, he found a square equal to the area of a region bounded by the parabola and a line parallel to the parabola's directrix. Why can a parabola be squared while a circle cannot?

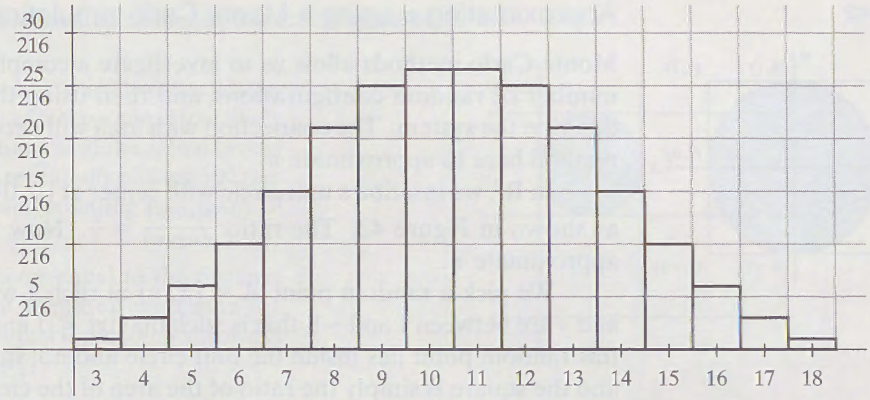
ANSWER TO QUESTION

$\overline{PQ}$ is the altitude to the hypotenuse of the right triangle whose vertices are Q and the endpoints of the diameter. So PQ is the geometric mean of the segments of the hypotenuse (Theorem 8.27).

10.1.6 Area as representing probability

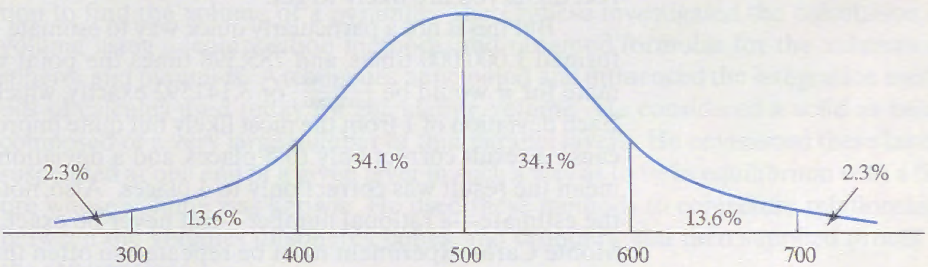
If a probability experiment has n possible mutually exclusive outcomes O_1 to O_n, and these outcomes have probabilities $p(O_1)$ to $p(O_n)$, then $\sum_{i=1}^{n} p(O_i) = 1$. The probabilities can be pictured in a histogram. Figure 40 shows a histogram for the probability of each possible sum when 3 fair dice, each with 1 to 6 on its faces, are tossed.

Figure 40



As the outcomes become more numerous, histograms like those in Figure 40 resemble curves. For instance, SAT scores were renormalized some years ago so that 500 is their mean and 100 is their standard deviation. Since any multiple of 10 from 200 to 800 is a possible score, there are 61 possible scores. Rather than draw a histogram with 61 bars, it is easier to draw a smooth (bell-shaped) curve connecting the tops of the bars, as in Figure 41. The area of the curve between $x = 445$ and $x = 605$ yields the probability that a randomly chosen test taker will score between 450 and 600, with the difference between the values of x and the scores due to the need to take into account rounding.

Figure 41



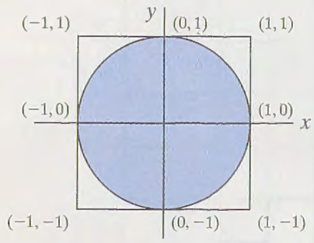
Using probability to determine area

In the preceding example, area represents probability. A class of techniques called *Monte Carlo methods* allow this idea to be turned around, for they use probability to determine area.

Monte Carlo methods are numerical methods that involve sampling from random numbers. They arose during World War II and owe their name to the similarities that can be made between statistical simulations and games of chance which are associated with the European gambling mecca Monte Carlo in the tiny country of Monaco. The Monte Carlo method can be used to simulate a Bingo game, simulate complex physical phenomena such as subnuclear processes in high-energy physics experiments, and study the flow of traffic in a city. There are also many applications of Monte Carlo methods in economics and computer science.

The idea of the Monte Carlo method is that a numerical problem in analytic form can be replaced by a problem in probability so that the numerical answer to the probability problem is the same as the numerical answer to the original problem. The new probability problem is solved using a computer program.

Figure 42



Approximating π using a Monte Carlo simulation

Monte Carlo methods allow us to investigate a complex system by sampling it in a number of random configurations, and then using the results of the sampling to describe the system. The connection with area will become apparent as we apply the method here to approximate π.

In $\mathbf{R}^2$, we inscribe a unit circle with center at $(0, 0)$ inside a square with vertices, as shown in Figure 42. The ratio $\frac{\alpha(\text{circle})}{\alpha(\text{square})} = \frac{\pi}{4}$. Now we use this ratio of areas to approximate π.

We pick a random point $A = (x, y)$ in such a way that the values for both x and y are between 1 and -1, that is, such that $|x| < 1$ and $|y| < 1$. The probability that this random point lies inside the unit circle and not in the space between the circle and the square is simply the ratio of the area of the circle to the area of the square.

$$P(x^2 + y^2 < 1) = \frac{\pi}{4}$$

Suppose we performed this experiment n times, and it turned out that x of those times produced a point inside the circle. Then we could estimate the probability to be $\frac{x}{n}$. As n approaches infinity, this probability estimate becomes arbitrarily close to $\frac{\pi}{4}$ and so we can write

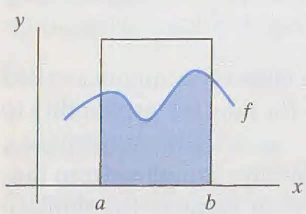
$$\lim_{n\to\infty} \frac{x}{n} = \frac{\pi}{4} \quad \text{or} \quad \pi = 4 \lim_{n\to\infty} \frac{x}{n}.$$

How precise is this formula for π? The precision (number of digits) depends on how many times you perform the experiment. The greater x and n are, the more correct digits you are likely to get.

But this is not a particularly quick way to estimate π. If the experiment were performed 1,000,000 times, and 785,398 times the point was inside the circle, the estimate for π would be $\frac{4\cdot 785{,}398}{1{,}000{,}000}$, or 3.141592 exactly, which is correct only to six places. Each deviation of 1 from the most likely but quite improbable result of 785,398 would cause a result correct only to 5 places, and a deviation of 10, still very likely, would mean the result was correct only to 4 places. Also, note that because π is irrational, the estimate—a rational number—can never be exact. When precision is needed, a Monte Carlo experiment must be repeated so often that computers are necessary.

For these reasons, Monte Carlo methods tend to be used when exact methods of calculation are unavailable or too unwieldy. For instance, if a function f had an equation for which the exact definite integral from $x = a$ to $x = b$ could not be calculated, then the graph of the function could be fit inside a rectangle with sides $x = a$ and $x = b$. By randomly selecting points in the rectangle and determining how many are under the graph of f, the definite integral could be estimated, as shown in Figure 43. Although this process is relatively tedious for calculating simple integrals, the Monte Carlo method is quite useful for calculating complicated integrals in n dimensions ($n \geq 3$), where other methods are relatively slow and expensive.

Figure 43



10.1.6 Problems

1. Write a scenario of a trip and a corresponding rate function and represent the total distance traveled as an area.

2. Determine the 16 probabilities in the histogram of Figure 40.

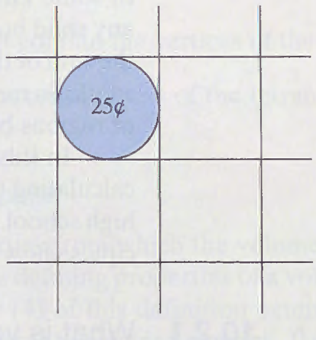
3. A general formula for a normal curve is $y = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}$ where μ is the mean and σ the standard deviation. What is a specific formula for the normal curve in Figure 41?

4. Use the idea of Figure 42 with at least 100 random points to estimate π.

5. Imagine an election that is even between two candidates. Suppose you conducted a survey of 500 people. How likely are your survey results to be within 2% of the actual (even) situation? Study this question by repeatedly picking 500 random numbers between 0 and 1 and counting how many of the 500 are less than 0.5.

6. Prove: If a quarter has diameter equal to the distance between sides in a square lattice of infinite extent (Figure 44), then the probability that a quarter that is randomly thrown at the lattice covers a lattice point is $\frac{\pi}{4}$.

Figure 44

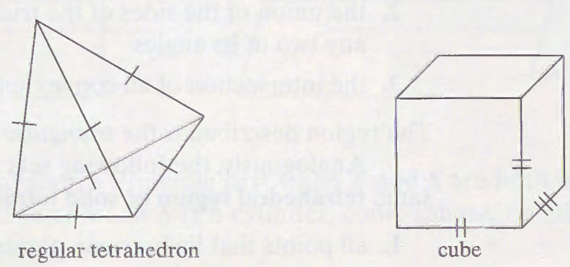


Unit 10.2 Volume

Ancient mathematicians needed to calculate volume to determine how much grain and other foodstuffs were stored in a particular location, or how much material would be needed to build structures or make other items. The Babylonians knew formulas for the volumes of boxes and, more generally, for the volume of a right prism with a trapezoidal base. They (incorrectly) found the volume of a truncated cone as the product of an altitude and half the sum of the areas of the bases. The *Rhind Papyrus* (which was a handbook for scribes) shows that the Egyptians computed the volumes of cylindrical granaries by multiplying the area of the circular base by the height. They were able to calculate the inclination of oblique planes, and used this calculation to find the volume of a pyramid. The Chinese investigated the calculation of volume using decomposition methods, and obtained formulas for the volumes of spheres and pyramids. Archimedes anticipated and influenced the integration methods of calculus used today for calculating volume. He considered a solid as being composed of a very large number of thin parallel layers. He envisioned these layers suspended at one end of a given lever in such a way as to be in equilibrium with a figure whose volume was known. He used these methods to conjecture relationships between the volumes of spheres, cones, and cylinders, and then supplied proofs of the relationships.

Can we develop a theory of volume by means of decomposition as we did with area? David Hilbert posed this question in 1900 when he encouraged mathematicians to investigate whether a definition of volume was possible for polyhedra analogous to that of area for polygons. Max Dehn (1878–1952) responded in that same year by showing that, although we can decompose two arbitrary polygonal regions of equal area into pairs of congruent triangles, it is not possible to decompose two arbitrary polyhedral regions of equal volume into pairs of congruent tetrahedra. He showed that decomposition is not always possible because figures as simple as a regular tetrahedron and a cube of equal volume (Figure 45) cannot be decomposed into congruent tetrahedral pieces.

Figure 45



As a consequence, like the areas of figures bounded by curves, infinite processes of some kind are needed to define volume for polyhedral regions. The volume of any solid bounded by plane polygons may be defined as the greatest lower bound of the sum of the volumes of nonoverlapping cubes, which together completely cover the solid. So methods of calculating volumes are analogous to those for calculating areas of regions bounded by curved lines.

In this unit, we investigate some figures of E^3, exploring different methods of calculating their volumes. We revisit some familiar formulas you probably learned in high school. Our goal is as it was with area, to derive these formulas from basic principles, show how they are related to each other, and illustrate them with applications.

10.2.1 What is volume?

We wish to define a volume function that will provide the means for obtaining the volumes of common 3-dimensional figures. You may wish at this time to refer back to the definition of area function found at the beginning of Section 10.1.1 and include the fourth part found at the start of Section 10.1.4.

Just as we think of area either as measuring the 2-dimensional space *occupied by* a region, or *contained in* the region's boundary, volume can be thought of either as a measure of the 3-dimensional space contained in a closed surface, or occupied by a solid that has that surface as a boundary. A **solid** is the set of points on a closed surface or in its interior.

Tetrahedrons

We began the study of area in Section 10.1.1 by considering the domain of the area function α to be the set of unions of triangular regions in E^2. This suggests that the domain of the volume function v might start from an analogous set in E^3. Such a set is the set of unions of solid *tetrahedrons* (an alternative plural is *tetrahedra*). Given four points not all in the same plane, a **tetrahedron** is the union of the four triangular regions (faces) determined by these points. That is, if the four given points are the noncoplanar points A, B, C, and D,

$$\text{tetrahedron } ABCD = \Delta ABC \cup \Delta ABD \cup \Delta ACD \cup \Delta BCD.$$

Such a tetrahedron is shown in Figure 46.

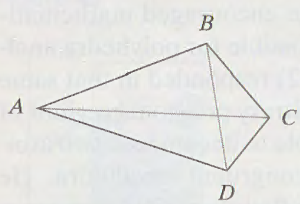


Figure 46

Tetrahedra are the spatial analogue of triangles. A triangle has three vertices, three sides, and three angles. A tetrahedron has four vertices, four faces, and four *solid angles*, one at each vertex. The solid angle at each vertex is the union of all points on and interior to the three plane angles of the tetrahedron at that vertex. Notice that these solid angles are not the twelve *plane angles* in the faces of the tetrahedron, nor are they the six *dihedral angles* that are formed by the union of the two half-planes that intersect at an edge of the tetrahedron.

The following sets associated with a triangle all describe the same region:

1. all points that lie between points on the sides of the triangle
2. the union of the sides of the triangle with the intersection of the interiors of any two of its angles
3. the intersection of all convex sets that contain the vertices of the triangle

The region described is the *triangular region* associated with the triangle.

Analogously, the following sets associated with a tetrahedron all describe the same **tetrahedral region** or **solid tetrahedron**.

1. all points that lie between points on the faces of the tetrahedron

2. the union of the faces of the tetrahedron with the intersection of the interiors of any two of its solid angles
3. the intersection of all convex sets that contain the vertices of the tetrahedron

The points in the tetrahedral region but not on any faces of the tetrahedron constitute the **interior** of the tetrahedron.

Defining properties of the volume function

In the following definition are four properties from which the volumes of solid figures can be derived. The first three of these defining properties of a volume function are analogous to those for area. Property (4) of this definition extends volume to figures that are not unions of a finite number of tetrahedral regions. It owes its name in the west to Bonaventura Cavalieri (1598–1647), a Jesuit priest who was a student of Galileo and who was the first western mathematician to realize its importance. But the first individuals to have used this principle to obtain volume seem to have been the Chinese mathematician Zu Chongzhi (429–500) and his son Zu Geng.

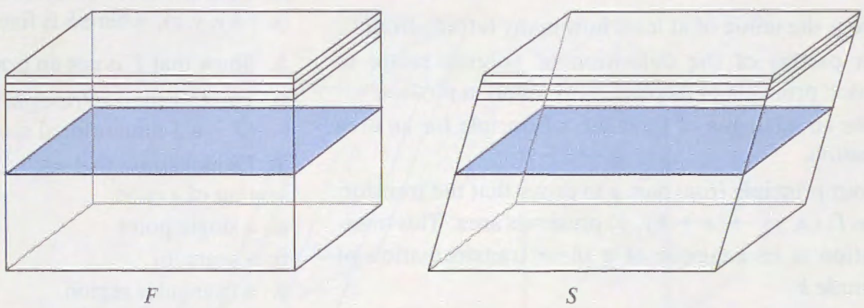
Definition Let F be the union of tetrahedral regions in E^3. A **volume function** v is a function that assigns to each such F a positive real number $v(F)$ such that:

1. If $F_1 \cong F_2$, then $v(F_1) = v(F_2)$. (Congruence property)
2. If the tetrahedral regions making up F_1 and F_2 have no interior points in common, then $v(F_1 \cup F_2) = v(F_1) + v(F_2)$. (Additive property)
3. If F is a cube with edge of length x, then $v(F) = x^3$. (Volume of cube)
4. If F and a solid S lie between parallel planes a and b, and for each plane c parallel to and between a and b, $c \cap S$ is a region whose area is known from the properties of the area function, and $\alpha(c \cap F) = \alpha(c \cap S)$, then $v(F) = v(S)$. (Cavalieri's Principle)

A **cross section** of a surface or solid is the intersection of a plane and the surface or solid. The intersections $c \cap F$ and $c \cap S$ in Cavalieri's Principle are cross sections.

To apply Cavalieri's Principle, we think of a 3-dimensional solid as being made up of parallel cross sections, as we do in calculus. If all the pairs of cross sections of two solids F and S made by parallel planes have the same area, then the solids have the same volume. You may think of a deck of cards F that has been slanted into the position S, as in Figure 47.

Figure 47



We will apply Cavalieri's Principle first when F and S are both unions of tetrahedral regions, and later when S is a cylinder, cone, sphere, or other figure with a curved boundary.

Volume is not area

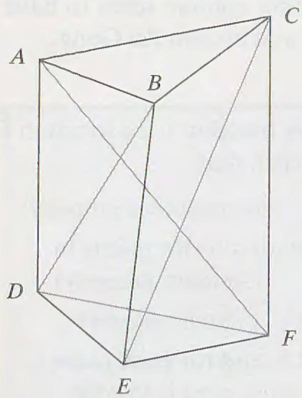
In Section 10.1.1, we began with the area formula for any square, yet we defined the area of any polygonal region in terms of triangular regions. We could do this because it is possible to decompose any polygonal region into triangular regions, which can then be arranged to form a square, as we saw in Section 10.1.5. We might say that we can "square" any polygon.

Can we approach the volume of polyhedra in an analogous way? To do so we need to address two issues, first, the decomposition of polyhedra, and second, the possibility of comparing polyhedra with equal volume in terms of dissections.

First let us address the notion of dissection. Is it possible to decompose any polyhedral region into tetrahedral regions analogous to the decomposition we have seen of polygonal regions into triangular regions?

It is easy to decompose any convex polyhedron P into tetrahedra. Triangulate all the faces of P and select a point A in the interior of P. Then the tetrahedra with vertex A and the three vertices of each triangle form a decomposition of P. If the polyhedron is not convex, like the one shown in Figure 48, where segments $\overline{AE}$, $\overline{BF}$, and $\overline{CD}$ are in the exterior of the polyhedron, then the situation is more complex. More than one point in the interior must be used for the decomposition.

Figure 48



We next address the issue of defining volume as we did area. Recall that we defined polygons to be equal in area if they possess dissections into corresponding congruent polygonal pieces. Can we define polyhedra to be equal in volume if they possess an analogous property? It turns out that we cannot make such a definition because in 1900 Max Dehn came up with a counterexample. Dehn showed that there is no way to decompose a regular tetrahedron and a cube of equal volume into equal numbers of tetrahedra that are congruent in pairs. We might say, as a result, that we cannot "cube" every polyhedron.

This limitation on the decomposition of polyhedra is why we cannot construct a theory of volumes analogous to that of area. However, Cavalieri's Principle does allow us to overcome this difficulty. We can find volume formulas for some solids not by decomposing them, but by equating them, cross section by cross section, with the volume of a known solid.

10.2.1 Problems

1. Prove that any convex solid must contain a tetrahedron.
2. A square pyramid is the union of at least how many tetrahedrons?
3. A cube is the union of at least how many tetrahedrons?
4. Which part(s) of the definition of volume relate to Archimedes' principle of displacement found in physics?
5. a. State an analogue of Cavalieri's Principle for an area function.

   b. Use your principle from part **a** to prove that the transformation $T: (x, y) \rightarrow (x + ky, y)$ preserves area. This transformation is an example of a *shear* transformation of magnitude k.

   c. Every parallelogram with base b and height h can be placed on a coordinate plane so that its vertices are $(0, 0)$, $(b, 0)$, (c, h), and $(c - b, h)$. What magnitude shear of part **b** maps this parallelogram onto a rectangle with area bh?

   d. Prove that with composition, the set of shears of the type shown in part **b** forms a group.
6. Consider the transformation T in $\mathbf{R}^3$ defined by $T(x, y, z) = (x + ky, y, z)$, where k is fixed.

   a. Show that T is not an isometry.

   b. Use Cavalieri's Principle to show that T preserves volume. (T is a 3-dimensional shear transformation.)
7. Demonstrate that each of the following is a possible cross section of a cube.

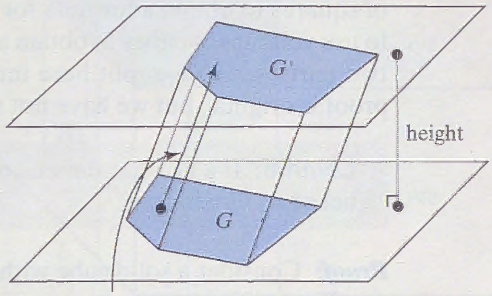
   a. a single point
   b. a segment
   c. a triangular region
   d. a rectangular region
   e. a square region
   f. a pentagonal region
   g. a hexagonal region

10.2.2 From cubes to polyhedra

Definitions of figures in space can be simplified by extending to E^3 some of the language we used earlier in E^2. Two figures α and β in E^3 are **congruent** if and only if there is a distance-preserving transformation T with $T(\alpha) = \beta$. One congruence transformation in E^3 is the translation associated with the 3-dimensional vector $\overrightarrow{AB}$. In R^3, if $\overrightarrow{AB} = (h, k, j)$, then the **translation associated with the vector (h, k, j)** has the formula $T(x, y, z) = (x + h, y + k, z + j)$. It is distinguished by the fact that all segments connecting points on a preimage with their images are parallel and of equal length.

Translations make it easier to define prisms and cylinders. Let G' be the translation image of a polygonal region G in E^3, where G' and G are not in the same plane. The **prismatic solid** with **bases** G and G' is the set of all points on any segment connecting a point of G with its translation image point on G' (see Figure 49). A **prism** is the boundary of this solid. The **altitude** or **height** of the prism is the (perpendicular) distance between the planes of G and G'.

Figure 49



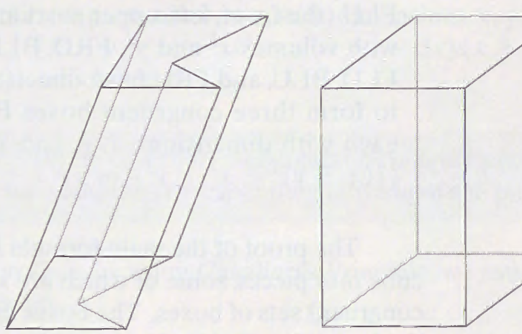
vector of translation T with T(G) = G'

Types of prisms

A prism is classified as **square, rectangular, triangular, quadrangular**, etc., as its base is a square, rectangle, triangle, quadrilateral, etc. In a prism, its bases are congruent (since translations are isometries) and the plane section formed by the prism's intersection with any plane parallel to and between its bases is a region congruent to the bases. Therefore, all these sections have the same area.

When the direction of the translation mapping one base to the other is perpendicular to the base planes, the prism is a **right prism**. Otherwise, it is called **oblique**. Figure 50a shows an oblique prism. Figure 50b shows a right prism.

Figure 50

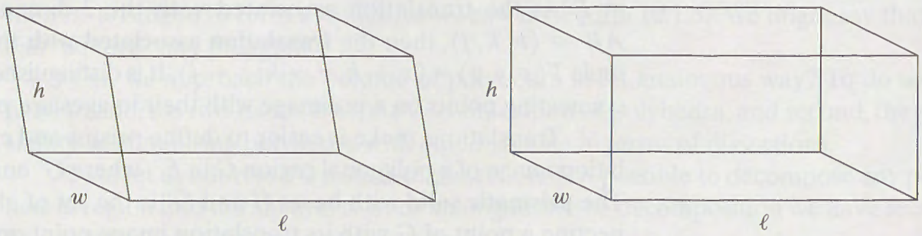


oblique prism (a)

right prism (b)

When a prism's base is a parallelogram, the prism is called a **parallelepiped** (Figure 51a). A right parallelepiped whose bases are rectangles is called a **right rectangular parallelepiped or box** (Figure 51b). A **cube** is a right rectangular parallelepiped whose faces and bases are squares.

Figure 51



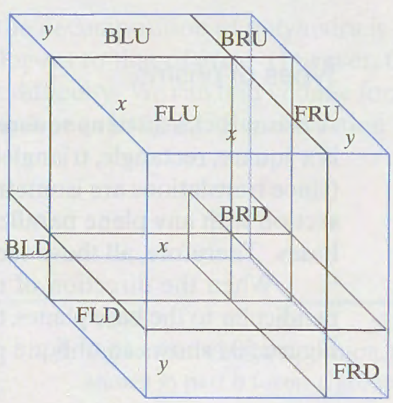
(a) parallelepiped (b) right rectangular parallelepiped

In developing the formulas for areas of polygonal regions, we first used areas of squares to obtain a formula for the area of any rectangle. The spatial analogue is to use volumes of cubes to obtain a formula for the volume of any box. The proof has two parts, which we split here into a lemma and the theorem. We doubt that the proof is original, but we have not seen it elsewhere.

Lemma: If a box has dimensions x, y, and $x + y$, then its volume is the product of its dimensions.

Proof: Consider a solid cube with edge $x + y$. It can be split into eight parts, shown in Figure 52.

Figure 52



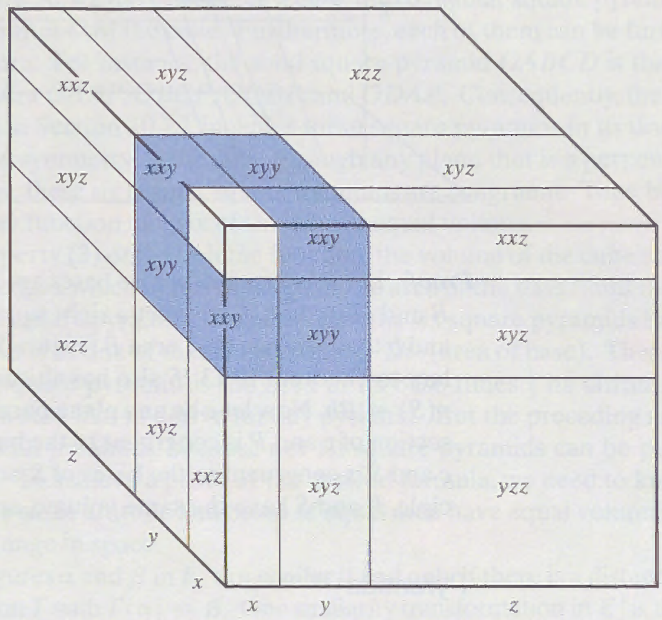
FLU (the front, left, upper part) and BRD (the back, right, lower part) are cubes with volumes x^3 and y^3. FRD, BLD, and BRU have dimensions x, y, and y, and FLD, BLU, and FRU have dimensions x, x, and y. These last six can be grouped to form three congruent boxes FRD ∪ FRU, BRU ∪ BLU, and BLD ∪ FLD, each with dimensions, x, y, and $x + y$. The volume of each of those boxes is $\frac{1}{3}((x + y)^3 - x^3 - y^3)$, or $xy(x + y)$. ∎

The proof of the main formula is similar to the proof of the lemma. We split a cube into pieces, some of which are smaller cubes. The rest we group together into congruent sets of boxes. The boxes in some of these sets have dimensions of form of the lemma, so their volumes are known. The remaining boxes have the dimensions we want, and their volumes can now be determined by subtracting the known volumes from the volume of the big cube.

Theorem 10.14 The volume of a right rectangular parallelepiped (box) with dimensions x, y, and z is xyz.

Proof: Consider a solid cube with edge $x + y + z$. It can be split into 27 parts, some of which can be seen in Figure 53.

Figure 53



Of the 27 parts, 3 are cubes with volumes x^3, y^3, and z^3. (Only the cube with volume x^3 can be seen in Figure 53.) Eighteen of the others can be grouped as follows:

Three have dimensions x, y, y and three have dimensions x, x, y. These join to form 3 boxes with dimensions $x, x + y$, and y, so by the lemma their joint volume is $3xy(x + y)$. (These are blue in Figure 53.)

Three have dimensions x, z, z and three have dimensions x, x, z. These join to form 3 boxes with dimensions $x, x + z$, and z, so by the lemma their joint volume is $3xz(x + z)$.

Three have dimensions y, z, z and three have dimensions y, y, z. These join to form 3 boxes with dimensions $y, y + z$, and z, so by the lemma their joint volume is $3yz(y + z)$.

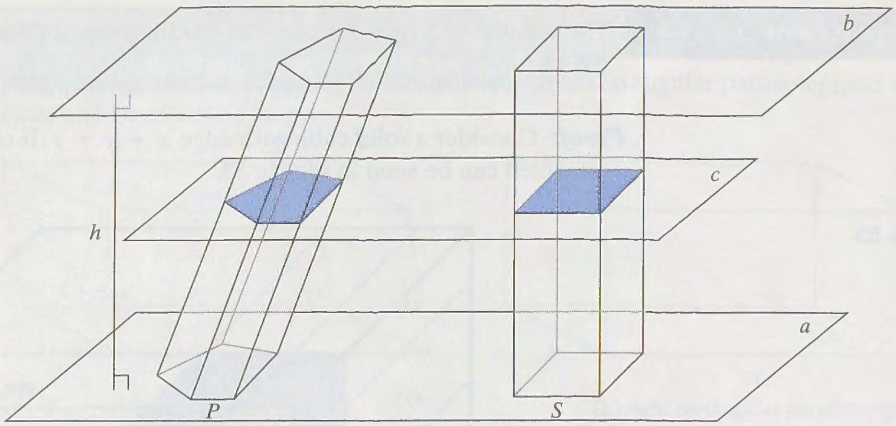
The remaining 6 parts are boxes with dimensions x, y, and z. The volume of each is $\frac{1}{6}((x + y + z)^3 - x^3 - y^3 - z^3 - 3xy(x + y) - 3xz(x + z) - 3yz(y + z))$, which is xyz. ∎

Corollary: The volume of a right rectangular parallelepiped is the product of the area of the base and the corresponding altitude of the parallelepiped.

From this corollary, by using Cavalieri's Principle, we can prove the following more general theorem.

Theorem 10.15 The volume of any prism with base area B and altitude h is Bh.

Figure 54



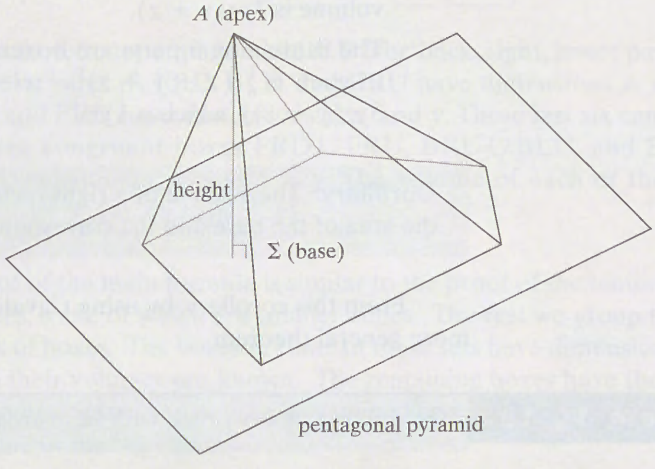
Proof: Let P be a prism whose bases are in planes a and b, and let P have base area B and altitude h. Construct a right square parallelepiped S with bases in planes a and b that also has base area B (Figure 54). This can be done because of the corollary to Theorem 10.13. S also has altitude h, so by the corollary to Theorem 10.14, $v(S) = Bh$. Now let c be any plane parallel to a or b and between them. The intersection of c and P is congruent to the bases of P, so has area B. The intersection of c and S is congruent to the bases of S, so also has area B. Thus, by Cavalieri's Principle, P and S have the same volume, and so $v(P) = Bh$. ∎

Pyramids

Although the previous arguments have obtained the volumes of prisms, in theory we have not yet established that prisms have a volume, for we have not shown a prismatic solid to be a union of tetrahedral regions. For this, we need to consider pyramids.

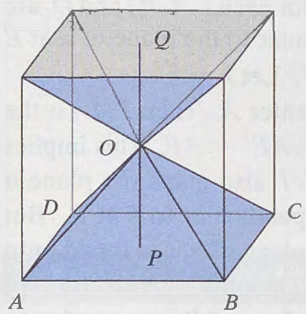
Given a polygonal region F in a plane, a **pyramidal solid** is the set of points on line segments connecting points of F (its **base**) with a point A (its **apex** or **vertex**) not in that plane. A **pyramid** is the boundary of this solid; it is union of F and the sets of points connecting the polygon boundary of F to A. The distance from the apex to the plane of the base is the **altitude** or **height** of the pyramid. Like prisms, pyramids are classified by their bases as triangular, quadrangular, pentagonal, etc. (Figure 55). From this definition, we can see that a triangular pyramid is a tetrahedron as defined in Section 10.2.1.

Figure 55



The word "pyramid" originates in the Greek word for the pyramids found in Egypt. These pyramids, and their counterparts built by the Mochica in Peru and the Maya in Central America, are *regular* square pyramids. A **regular pyramid** is a pyramid whose base F is a regular polygon and in which the segment connecting the apex to the center of F is perpendicular to the plane of F. That is, in a regular pyramid, if the base is horizontal, the apex is directly above the center of the base.

Figure 56



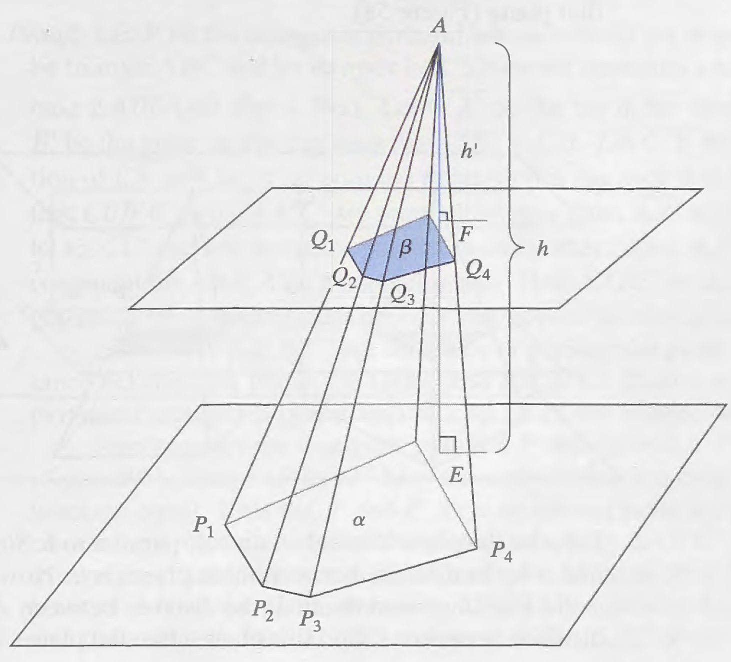
In Figure 56, we have dissected a cube into 6 regular square pyramids, each with apex at the center O of the cube. Furthermore, each of them can be further dissected into tetrahedra. For instance, the solid square pyramid $OABCD$ is the union of the solid tetrahedra $OABP$, $OBCP$, $OCDP$, and $ODAP$. Consequently, the volume function defined in Section 10.2.1 includes these square pyramids in its domain. Due to the reflection symmetry of the cube through any plane that is a perpendicular bisector of an edge, these six regular square pyramids are congruent. Thus, by property (1) of the volume function, the six of them have equal volume.

By property (3) of the volume function, the volume of the cube equals the cube of one of its edges, which in this case equals the area of the base times twice the height of any of the square pyramids. Since each of the six square pyramids have equal volume, the volume of one of them must equal $\frac{1}{6} \cdot 2h \cdot$ (area of base). Therefore, the volume of the square pyramid is the area of its base times $\frac{1}{3}$ its altitude. This is the formula we would like to derive for any pyramid. But the preceding argument does not apply to all pyramids, because not all square pyramids can be put together to form a cube. To achieve a proof of the desired formula, we need to know that pyramids with the same altitude and bases of equal area have equal volumes. For this, we use a size change in space.

Two figures α and β in E^3 are **similar** if and only if there is a distance-multiplying transformation T with $T(\alpha) = \beta$. One similarity transformation in E^3 is the **size change with center O and magnitude k > 0**. As in the two-dimensional case, the image of any point P under this size change is the point P' such that P' is on ray $\overrightarrow{OP}$ and $\frac{OP'}{OP} = k$.

Now we prove a lemma about the relationship between the area of a cross section of a pyramid parallel to its base and the area of the base itself.

Figure 57



Lemma: Let P be a solid pyramid with altitude h and base α that has area B. Let Q be a pyramid with the same apex as P, and such that its base β is a section of P contained in a plane that is parallel to the plane of α. Let the area of the base of Q be B' and the altitude of Q be h'. Then $\frac{B'}{B} = \left(\frac{h'}{h}\right)^2$.

Proof: Let α be the polygon $P_1P_2 \ldots P_n$ and let A be the common apex of P and Q (Figure 57). Let β be the polygon $Q_1Q_2 \ldots Q_n$, so that for each i, A, P_i, and Q_i are collinear. Let ℓ be the line through A that is perpendicular to the plane of α at E and to the plane of β at F. Then $AE = h$ and $AF = h'$. Let $k = \frac{h'}{h}$.

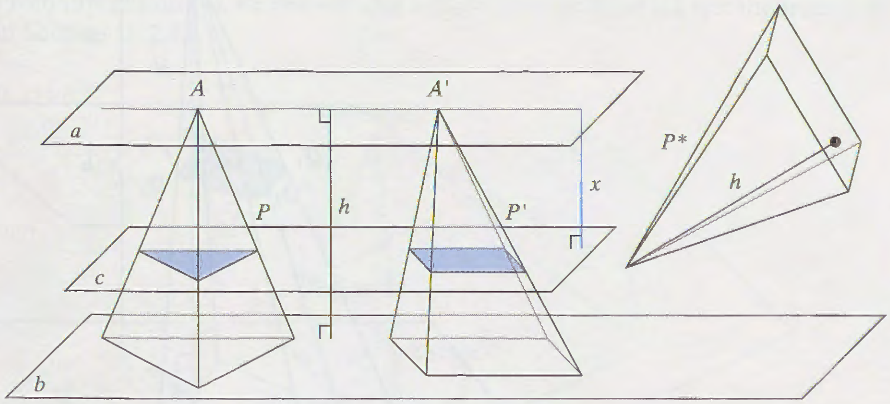
Let T be a size change in space of magnitude k, center A. Then $T(E)$ is the point E' on $\overrightarrow{AE}$ such that $AE' = kAE$. Since $k = \frac{AF}{AE}$, $AE' = AF$. This implies that $E' = F$ since F is also on $\overrightarrow{AE}$. Since $T(E) = F$, T also maps the plane α (which is perpendicular to ℓ at E) onto the plane perpendicular to ℓ at F. But this is the plane of β. Therefore, every point X of the plane of α is mapped onto the point in the plane of β at which $\overrightarrow{AX}$ intersects the plane. Thus T maps $P_1, P_2, \ldots, P_n$ to $Q_1, Q_2, \ldots, Q_n$, respectively. And so the base of P is mapped onto the base of Q. These bases are similar and k is the ratio of similitude. The area of $Q_1Q_2 \ldots Q_n = k^2$ times the area of $P_1P_2 \ldots P_n$. Thus $B' = k^2B$, and the lemma follows. ∎

We now use Cavalieri's Principle to show that the volume of a pyramid depends only on its base area and altitude.

Theorem 10.16 If two pyramids have the same altitude and same base areas, then they have the same volume.

Proof: Suppose P is any pyramid with a base in plane b. Any other pyramid P^* with the same height h and base area as P is congruent to a pyramid P' with height h whose base is in the same plane as the base of P and that lies on the same side of that plane (Figure 58).

Figure 58



Let a be the plane through A and A' parallel to b. Since the altitude of each pyramid is h, the distance between these planes is h. Now let c be any plane parallel to a and b and between them. If the distance between A and this plane is x, so is the distance between A' and this plane. (Parallel planes are everywhere equidistant.)

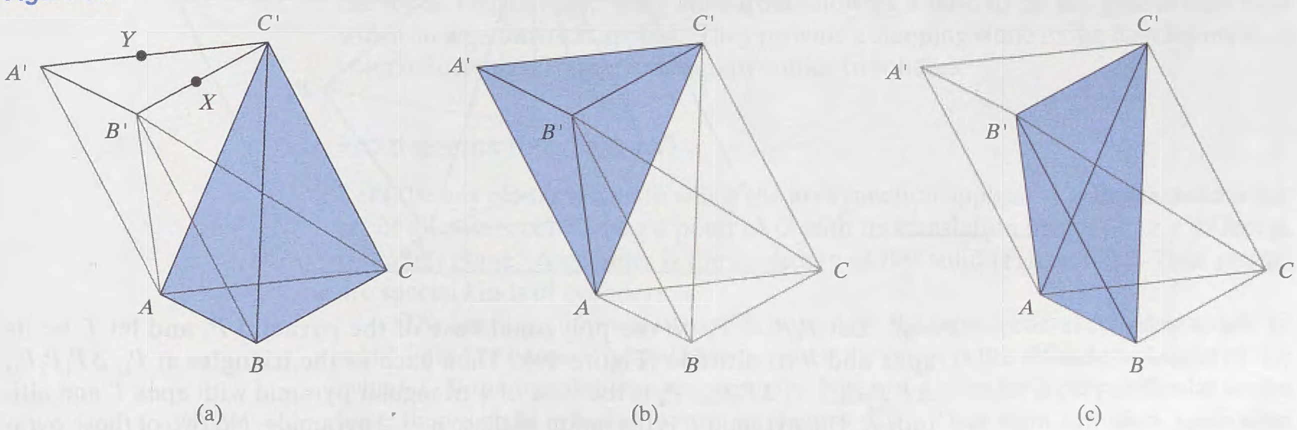
If B is the area of the base in plane b, then, by the lemma, the area of the intersection of c and P has area $\left(\frac{x}{h}\right)^2 B$. By the same argument, the area of the intersection of c and P' is $\left(\frac{x}{h}\right)^2 B$. Since the areas of all parallel cross sections are equal, we can conclude by Cavalieri's Principle that the volumes of P and P' are equal. ∎

From tetrahedra to pyramids

In the definitions of the area function (Section 10.1.1) and the volume function (Section 10.2.1), triangles and tetrahedra play analogous roles. Now we shall see that they also play the analogous roles in the derivation of formulas. That is, just as we put triangular regions together to get formulas for the areas of some polygons, we now put volumes of tetrahedral regions together to obtain formulas for the volumes of some pyramids. We first show that the volume of a tetrahedron is equal to one-third the area of its base times its altitude. We then generalize the formula to all pyramids.

Theorem 10.17 The volume of a tetrahedron with base area B and altitude h is $\frac{1}{3}Bh$.

Figure 59



Proof: Let P be the triangular pyramid whose volume we want. Let the base of P be triangle ABC and let its apex be C'. Now we construct a triangular prism with base ΔABC (see Figure 59a). Let $\overrightarrow{C'X}$ be the ray in the direction of $\overrightarrow{CB}$ and let B' be the point on this ray such that $C'B' = CB$. Let $\overrightarrow{C'Y}$ be the ray in the direction of $\overrightarrow{CA}$ and let A' denote the point on this ray such that $C'A' = CA$. Notice that $CBB'C'$ and $CAA'C'$ are parallelograms. Thus, $\overline{AA'}$ and $\overline{BB'}$ are both parallel to $\overline{CC'}$ and are therefore parallel to each other. Also $\overline{AA'}$, $\overline{CC'}$, and $\overline{BB'}$ are congruent, so $ABB'A'$ is a parallelogram. Thus $\Delta ABC \cong \Delta A'B'C'$ by SSS Congruence (corresponding sides are opposite sides of parallelograms).

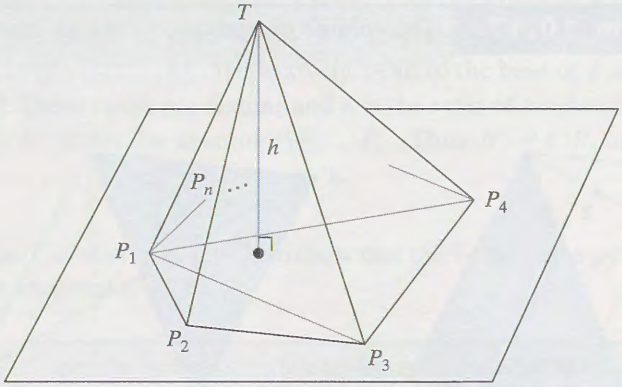
ΔABC and $\Delta A'B'C'$ are the bases of a triangular prism Q. Let h be the distance between the planes of ΔABC and $\Delta A'B'C'$. Then h is the altitude of both pyramid P and prism Q and, by Theorem 10.15, the volume of $Q = h \cdot \alpha(\Delta ABC)$.

Now consider the triangular pyramid P' with base $\Delta A'B'C'$ and with apex A (Figure 59b). Because P and P' have the same altitude and congruent bases, their volumes are equal. Note that P and P' have no interior point in common as their interiors are on opposite sides of the plane containing $\Delta AB'C'$. Next let P'' be the triangular pyramid with base $\Delta BC'B'$ and apex A (Figure 59c). Let h' denote the distance of A from the plane containing $\Delta BCB'$. We can also describe the pyramid P as a pyramid with base $\Delta C'BC$ and apex A. But $\Delta BC'B' \cong \Delta C'BC$ by SSS, and therefore these triangles have the same area. From this, using Cavalieri's Principle, it follows that the volume of P is equal to the volume of P''. Again, P and P'' have no interior point in common as they are on opposite sides of the plane of $\Delta ABC'$. Nor do P' and P'' have a point in common, since they are on opposite sides of the plane of $\Delta AB'C'$.

Since the triangular prism Q is the union of P, P', and P'', and no two of these tetrahedra have an interior point in common, the volume of the prism Q is the sum of the volumes of the tetrahedra. Thus the volume of Q is three times the volume of P, and the theorem follows. ┘

Theorem 10.18 The volume of a pyramid with base area B and altitude h is $\frac{1}{3}Bh$.

Figure 60



Proof: Let $P_1P_2 \dots P_n$ be the polygonal base of the pyramid P, and let T be its apex and h its altitude (Figure 60). Then each of the triangles at P_1: $\Delta P_1P_2P_3$, $\Delta P_1P_3P_4, \dots, \Delta P_1P_{n-1}P_n$ is the base of a triangular pyramid with apex T and altitude h. The pyramid P is the union of these $n-2$ pyramids. No two of these pyramids have an interior point in common. Therefore, the volume of P is simply the sum of the volumes of these triangular pyramids, and the theorem follows. ┘

10.2.2 Problems

1. In 1999, the human population of Earth passed 6 billion. Could all people on Earth then fit into a cube with edges 1 mile long?

2. a. The diagonals of the faces of a box have lengths 3, 4, and 6. What is the volume of the box?

   b. Find a formula for the volume of a box whose face diagonals have lengths a, b, and c.

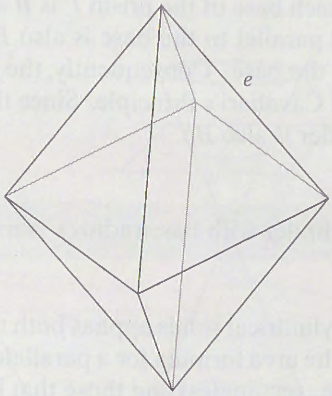
3. Prove the following theorem due to Legendre: Given a parallelepiped P, a rectangular parallelepiped can be constructed that has the same volume, same height, and same base area as P.

4. A **median of a tetrahedron** $ABCD$ is a segment from one vertex (say A) of a tetrahedron to the centroid of the opposite face (ΔBCD). (Recall that the centroid of a triangle is the point of concurrency of the medians of the triangle.) Prove that any plane passing through a median of a tetrahedron and containing a second vertex of the tetrahedron bisects the volume of the tetrahedron.

5. In Figure 56, if the volume of tetrahedron $OABC$ is V, what is the volume of the cube?

6. A regular octahedron is an 8-sided polyhedron whose faces are all equilateral triangles, as shown in Figure 61. What is the volume of a regular octahedron with edge e?

Figure 61



7. The pyramid of Khufu at Ghiza outside of Cairo is the largest of the Egyptian pyramids. It is a square pyramid with sides about 230 m long and had an original height of 147 m. The height is now about 137 m due to the loss of its outer stones.
   a. Estimate its original volume.
   b. Estimate its current volume.
   c. The pyramid contains about 2.3×10^6 stone blocks, each weighing about 2.75 tons. What is the volume of each block (assuming they are the same size)?

8. Is there a 3-dimensional analogue to a Reuleaux triangle (See Problem 6, Section 10.1.1)? That is, is there a solid R that is not a sphere such that, for any pair of parallel planes tangent to R, the distance between the plane is constant? If so, describe R. If not, explain why not.

10.2.3 From polyhedra to spheres

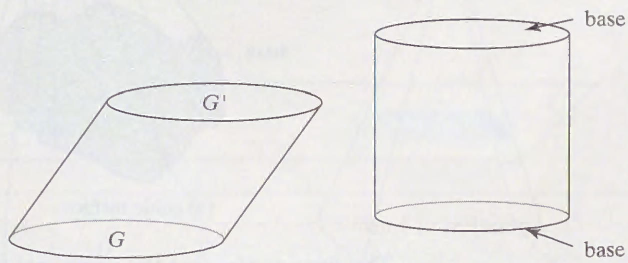
Cylindric solids and *conic solids* are generalizations of prismatic solids and pyramidal solids, respectively. They arise from allowing a base to be any planar region to which an area function applies. They provide a stepping stone in the development of volume formulas from prisms and pyramids to spheres.

From prisms to cylinders

Let G be any planar region to which the area function applies. A **cylindric solid** is the set of all points connecting a point of G with its translation image G' in a different (parallel) plane. A **cylinder** is the boundary of this solid (Figure 62a). Thus prisms are special kinds of cylinders.

The language of prismatic solids is used for the more general cylindric solids. G and G' are the **bases**. The distance between the bases is the **altitude** or **height** of the cylinder. If the translation connecting the bases of a cylinder is perpendicular to the base, then a **right cylinder** is formed (Figure 62b). Thus right cylinders generalize right prisms. Otherwise, the cylinder is called **oblique**. When the bases G and G' are circles, then a **circular cylinder** is formed.

Figure 62



(a) oblique cylinder (b) right cylinder

The volume of any cylinder can be found if the area of its base is known.

Theorem 10.19 The volume of a cylinder with base area B and altitude h is Bh.

Proof: Let C be a cylinder with bases G_1 and G_2, height h, and let $B = \alpha(G_1) = \alpha(G_2)$. Because each cross section of the cylinder parallel to the base is congruent to the base, any cross section has area B. Now let P be a prism with rectangular bases of dimensions 1 by B in the same planes as G_1 and G_2. (We do not show a figure; you should draw one.) The area of each base of the prism P is B and its height is h. The area of any cross section of P parallel to the base is also B because all parallel cross sections are congruent to the base. Consequently, the prism P and cylinder C satisfy the given conditions of Cavalieri's Principle. Since the volume of the prism is Bh, the volume of the cylinder is also Bh. ∎

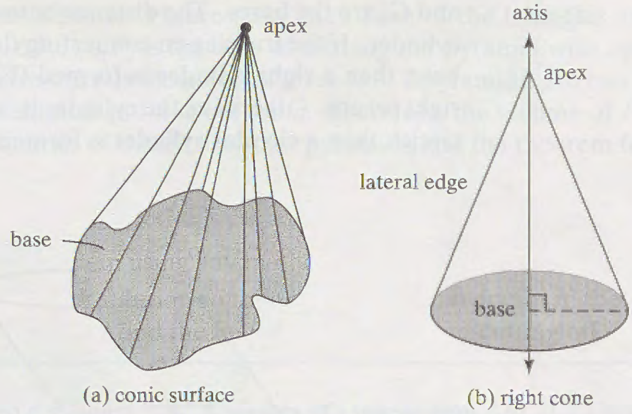
Corollary: The volume of a circular cylinder with base radius r is $\pi r^2 h$.

Notice that the volume formula for cylindrical solids applies both to right cylinders and oblique cylinders. This is akin to the area formula for a parallelogram applying both to right-angled parallelograms (i.e., rectangles) and those that have no right angles. This tends to violate many people's intuition about the volumes of these figures, which is affected by the different boundaries—surface area for the cylinders and perimeter for the parallelograms.

From pyramids to cones

The same principles that generalize prisms to cylinders also generalize pyramids to cones. Given a connected region G in a plane, a **conic solid** is the set of all points on line segments connecting a point of G (its **base**) with a single point (its **apex**) in a different plane (Figure 63a). The conic solids of main interest are those in which G is a circle. A **circular cone** is the boundary of a conic solid whose base is a circle. If the segment connecting the apex of the cone to the center of the base is perpendicular to the base, then a **right circular cone** is formed (Figure 63b). Right circular cones correspond to regular pyramids.

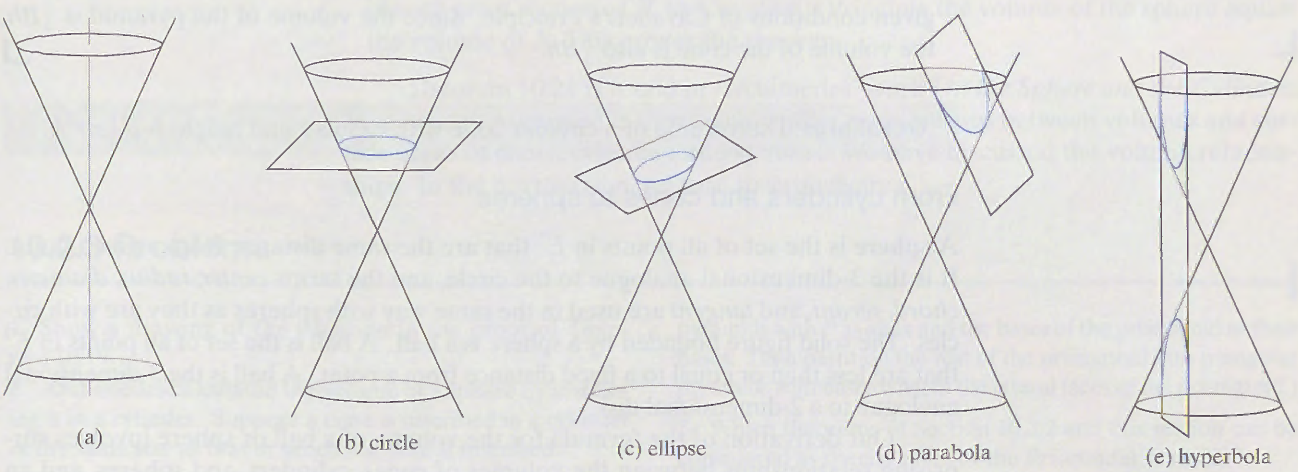
Figure 63



The *lines* containing the segments that join the apex to the points of the circular base in a right circular cone form a figure of infinite extent that is open on each side of the apex. This figure is called a **two-napped cone** (Figure 64a), and from it the conic sections arise (Figures 64b–64e). This provides another way of arriving at a right circular cone. Form a two-napped cone by rotating one of two intersecting lines in space about the other line at the point of intersection. Then cut off one of the nappes by a plane perpendicular to the line that is fixed to form a right circular cone.

This gives the word "cone" two closely related but different meanings, one meaning used when studying area and volume, the second used when studying the conic sections.

Figure 64

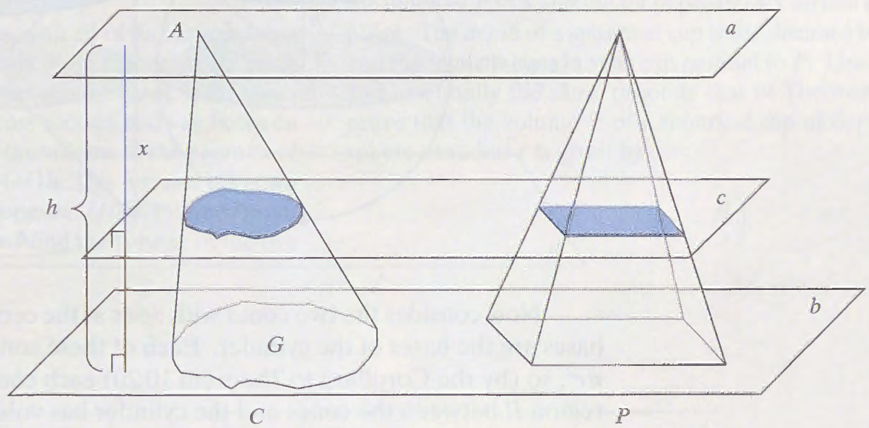


The derivation of a volume of a cone is slightly more complex than for a cylinder but also involves Cavalieri's Principle.

Theorem 10.20 The volume of a cone with base area B and altitude h is $\frac{1}{3}Bh$.

Proof: We begin as in the proof of Theorem 10.19. Let C be a cone with apex A, base G in plane b, and height h, and let $B = \alpha(G)$. Now let P be a pyramid with a rectangular base of dimensions 1 by B in the plane b, and apex in the plane a through A parallel to the plane of the base (Figure 65). Then the height of P is also h. Let c be a plane between a and b at a distance x from a, creating a cross section of P. (This argument should by now be rather familiar.) Because the cross section can be thought of as the image of the base under a size change with center A and magnitude $\frac{x}{h}$, the area of the cross section is $B\left(\frac{x}{h}\right)^2$.

Figure 65



The intersection of the plane c with the cone C is a cross-section that is also similar to the base, being the image of the base under a size change with center at the cone's apex and magnitude $\frac{x}{h}$. The cone's base area B is the sum of areas of triangles or the greatest upper bound of such sums. The cross-section's area is the sum of areas of the size change images of these triangles. Because each triangle in calculating the cone's cross-sectional area has area $\left(\frac{x}{h}\right)^2$ times the corresponding triangle in the base, the cross-section has area $B\left(\frac{x}{h}\right)^2$. Consequently, the pyramid P and cone C satisfy the given conditions of Cavalieri's Principle. Since the volume of the pyramid is $\frac{1}{3}Bh$, the volume of the cone is also $\frac{1}{3}Bh$. ┘

Corollary: The volume of a circular cone with radius r and height h is $\frac{1}{3}\pi r^2 h$.

From cylinders and cones to spheres

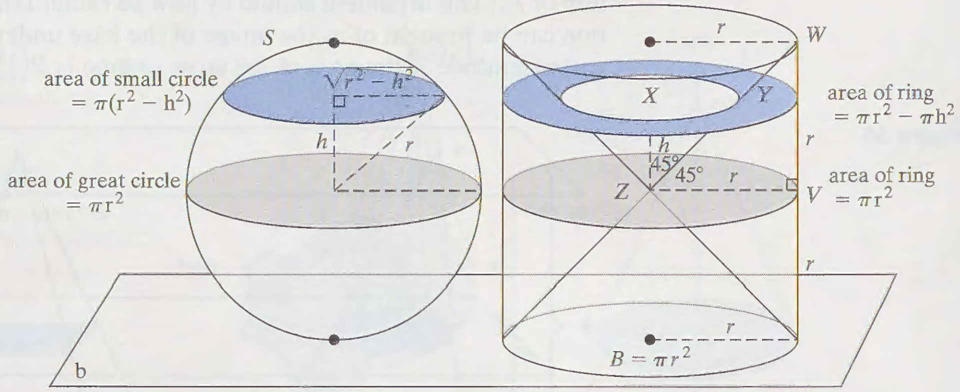
A **sphere** is the set of all points in E^3 that are the same distance from a given point. It is the 3-dimensional analogue to the circle, and the terms *center, radius, diameter, chord, secant,* and *tangent* are used in the same way with spheres as they are with circles. The solid figure bounded by a sphere is a **ball**. A ball is the set of all points in E^3 that are less than or equal to a fixed distance from a point. A ball is the 3-dimensional analogue to a 2-dimensional disk.

Our derivation of the formula for the volume of a ball or sphere involves surprising relationships between the volumes of cones, cylinders, and spheres, and an elegant application of Cavalieri's Principle.

Theorem 10.21 The volume of a sphere with radius r is $\frac{4}{3}\pi r^3$.

Proof: Let S be a sphere with radius r, and let a and b be planes tangent to the sphere at the endpoints of a diameter of S. To use Cavalieri's Principle, we construct a right circular cylinder with radius r and bases in the planes a and b. This cylinder has height $2r$, so (by the Corollary to Theorem 10.19) its volume is $2r \cdot \pi r^2$, or $2\pi r^3$. In Figure 66 we have shown the sphere, cylinder, and plane b. The plane a is not shown because it would hide helpful details.

Figure 66



Now consider the two cones with apex at the center of the cylinder and whose bases are the bases of the cylinder. Each of these cones has height r and base area πr^2, so (by the Corollary to Theorem 10.20) each cone has volume $\frac{1}{3}\pi r^3$. Thus the region R between the cones and the cylinder has volume $2\pi r^3 - 2 \cdot \frac{1}{3}\pi r^3$, or $\frac{4}{3}\pi r^3$.

We now show that the volume of the sphere equals the volume of R. Consider a cross section parallel to a and b. If the cross section contains the center of the sphere, then its area is πr^2 in both the sphere and the cylinder. If the cross section of the sphere is at a distance h from the center, then, as Figure 66 shows, its area is $\pi(r^2 - h^2)$. The corresponding cross section of R is a ring. The outside radius of the ring is r. Notice that $WV = VZ = r$, so $m\angle WZV = 45°$. Consequently, $m\angle WZX = 45°$. Thus $XY = XZ$ and the inner radius of the ring is h. Consequently, the area of the ring is $\pi r^2 - \pi h^2$, the same as the area of the cross section of the sphere. Because the area of each cross section of the sphere equals the area of each cross section of R, by Cavalieri's Principle the volume of the sphere equals the volume of R. This proves the theorem. ⌟

Theorem 10.21 is found in Archimedes' work *On the Sphere and the Cylinder.* In that work Archimedes shows many other relationships between volumes and surface areas of cones, cylinders, and spheres. We have discussed the volume relationships. In the next section, we turn to surface area.

10.2.3 Problems

1. Show a drawing of the situation in the proof of Theorem 10.19.

2. Archimedes calculated the volume of a sphere by inscribing it in a cylinder. Suppose a cone is inscribed in a cylinder of the same size as that in which a sphere is inscribed.

a. What is the ratio of the volume of the cone to the volume of the cylinder?

b. What is the ratio of the volume of the sphere to that of the cone?

3. What common figure has a volume equal to $\frac{2}{3}hB$, where h is its height and B the area of its base?

4. Rain gauges are often made in the shape of cones. Figure 67 shows two such cones of the same height H but different radii r_1 and r_2.

a. Show that if these rain gauges are placed in nearby locations in a rain storm, the water will reach the *same height* h in each of them, independent of their diameters.

b. How does this height h vary with the amount d of rainfall?

5. Criticize the following explanation: To say that the volume of a sphere is 400 cubic feet means that the sphere contains 400 cubic feet, or that the amount of space is the same as in a 400-foot cube.

6. A **prismatoid** is a polyhedron with all of its vertices lying in two parallel planes. The faces in those planes are its bases and its altitude h is the distance between the bases. If the bases have areas B and B' and the cross section midway between the bases has area M, show that the volume of the prismatoid is given by $V = \left(\frac{1}{6}\right)(B + B' + 4M)h$. This formula is known as the **Prismoid** or **Prismoidal Formula**. (*Hint:* Pick any point P on the middle cross section, and find the volume of the two pyramids with P as apex and the bases of the prismatoid as their bases. Then partition the rest of the prismatoid into triangular pyramids with bases lying in the lateral faces of the prismatoid.)

7. Which theorems of Section 10.2.2 and this section can be considered as special cases of the Prismoidal Formula?

8. Find the volume of the largest right circular cone in which a sphere of radius 5 can be inscribed.

9. Derive the formula for the volume of a circular cone using calculus. (Consider the cone as a solid formed by rotating a right triangle in space about one of its legs.)

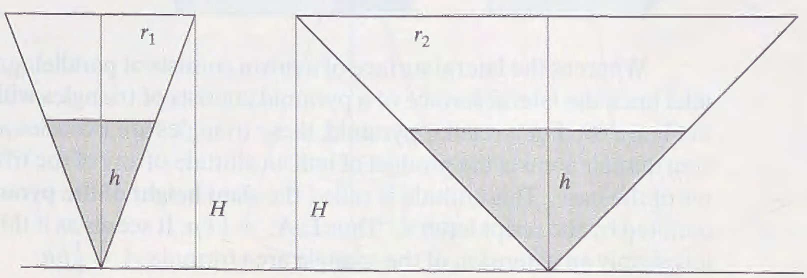
10. Derive the formula for the volume of a sphere using calculus. (Consider the sphere as a solid formed by rotating a semicircle about one of its diameters.)

11. Prove that there is exactly one sphere that contains the vertices of a tetrahedron. (*Hint:* Let A, B, C, and D be the vertices, and let E be the circumcenter of triangle ABC. Prove that if X belongs to a line ℓ that is perpendicular to the plane of ΔABC at E, then $AX = BX = CX$. Then prove that if Y is a point equidistant from A, B, and C, then Y belongs to ℓ.)

*12. Let P be a plane intersecting a ball S. A **spherical cap** is the union of $P \cap S$ and the set of points of S on one side of the plane. The *depth* of a spherical cap is the distance between P and the plane tangent to the cap parallel to P. Use Figure 66 and essentially the same proof as that of Theorem 10.21 to prove that the volume V of a spherical cap of depth h for a sphere of radius r is given by

$$V = \pi h^2\left(r - \frac{h}{3}\right).$$

Figure 67



Unit 10.3 Relationships among Area, Volume, and Dimension

While perimeter, area, and volume measure different aspects of figures in different dimensions, and while knowing one of these seldom determines any other, these measures are related in many ways. For instance, every area formula involves the product of two lengths, and every volume formula involves the product of three lengths. Also, a rectangle with a given area can have as large a perimeter as one wants, but no smaller than the perimeter of a square with that area. In this unit we discuss a variety of other interesting relationships among length, area, and volume.

10.3.1 Surface area

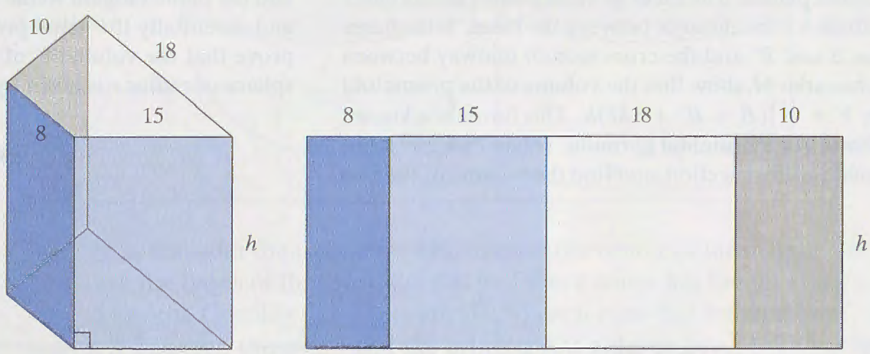
The surface of any prism, pyramid, polyhedron, cylinder, or cone can be folded or rolled onto a plane. As a consequence, the surface areas of these 3-dimensional figures are found in the same way that 2-dimensional areas are found. When there are new formulas, it is only because the arrangement of the surface on the plane has regularity that allows the area to be calculated given certain dimensions of the original 3-dimensional surface. However, the use of an abbreviation such as S.A. for surface area misleads some students to believe that surface area is as different from area as volume is. The sphere is the only commonly studied solid whose surface cannot be rolled or folded to be a plane figure. Other common solids with this characteristic are ellipsoids and tori (doughnut-shaped solids).

Surface areas of prisms and pyramids

The part of the surface of a cylindric or conic solid that is not the base is called the **lateral surface** of the solid. The lateral surface of any prism can be unfolded to form a union of parallelograms. The total area of these parallelograms is the **lateral surface area (L.A.)** of the prism. The bases can then be attached at opposite sides of one of the parallelograms. The resulting figure is a **net** for the prism, and its area is the **total surface area (S.A.)** of the prism.

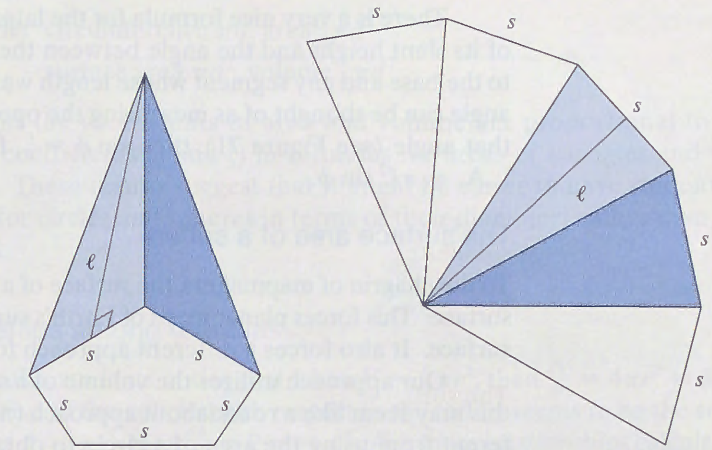
In school mathematics, surface areas of prisms are most often found for right prisms. Then the parallelograms of the net are rectangles and the lateral area is the product of the height of the prism and the perimeter of its base. An example is shown in Figure 68.

Figure 68



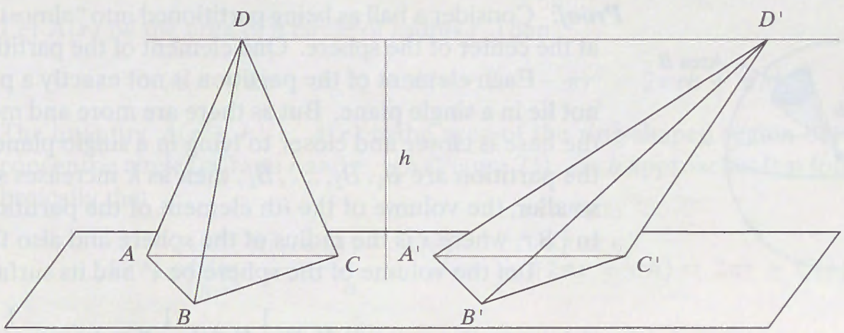
Whereas the lateral surface of a prism consists of parallelograms between two parallel lines, the lateral surface of a pyramid consists of triangles with a common vertex, as in Figure 69. For a regular pyramid, these triangles are isosceles and congruent, and the sum of their areas is the product of half an altitude of any of the triangles and the perimeter of the base. This altitude is called the **slant height of the pyramid** and is customarily denoted by the script letter ℓ. Thus L.A. $= \frac{1}{2}\ell p$. It seems as if this is a new formula, but it is simply an extension of the triangle area formula $A = \frac{1}{2}bh$.

Figure 69



In Figure 70 we show two pyramids with congruent triangular bases ABC and $A'B'C'$ and the same height h. Notice that, whereas the volumes of the pyramids are equal, their surfaces areas are not equal. This is a 3-dimensional analogue to the fact that triangles with equal bases and altitudes can have different perimeters.

Figure 70

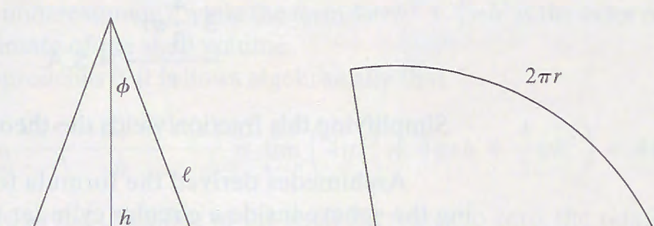


Surface areas of cylinders and cones

Unrolling the label from a soup can mimics the process used to determine the lateral area of a cylinder. When unrolled, the lateral surface of a right cylinder is a rectangle whose height is the height of the cylinder and whose width is the circumference of the base.

Unrolling a right circular cone is a little more interesting. The lateral surface is a sector of a circle. By thinking of the area of the sector as a limit of the sum of areas of triangles, the formula L.A. $= \frac{1}{2}\ell p$ for pyramids can be adapted. Here ℓ is the **slant height of the cone**, the length of any segment joining the apex to any point on the circle that is the boundary of the base; p is the perimeter of the base, or $2\pi r$, if the radius of the base is r. So one formula for the lateral area of a right circular cone is L.A. $= \pi r\ell$.

Figure 71



There is a very nice formula for the lateral area of a right circular cone in terms of its slant height and the angle between the segment from the apex perpendicular to the base and any segment whose length was used to measure the slant height. This angle can be thought of as measuring the opening of the cone. If ϕ is the measure of that angle (see Figure 71), then $\sin\phi = \frac{r}{\ell}$, from which $r = \ell \sin\phi$. Consequently, L.A. $= \pi\ell^2 \sin\phi$.

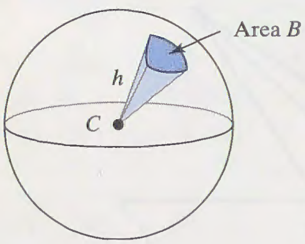
The surface area of a sphere

To the chagrin of mapmakers, the surface of a sphere cannot be unrolled onto a plane surface. This forces planar maps of Earth's surface to distort some aspect of the actual surface. It also forces a different approach for obtaining the surface area.

Our approach utilizes the volume of a sphere to obtain its surface area. While this may seem like a roundabout approach (no pun intended), it is not that much different from using the area of a circle to obtain a value for π and thus to obtain its circumference.

Theorem 10.22 The surface area of a sphere with radius r is $4\pi r^2$.

Figure 72



Proof: Consider a ball as being partitioned into "almost pyramids", each with an apex at the center of the sphere. One element of the partition is shown in Figure 72.

Each element of the partition is not exactly a pyramid because its base does not lie in a single plane. But as there are more and more elements of the partition, the base is closer and closer to lying in a single plane. If the areas of the bases of the partition are $B_1, B_2, \ldots, B_k$, then as k increases so that the largest B_k is made smaller, the volume of the ith element of the partition becomes closer and closer to $\frac{1}{3}B_i r$, where r is the radius of the sphere and also the height of the pyramid.

Let the volume of the sphere be V and its surface area be S.A.

$$\begin{aligned} V &\approx \frac{1}{3}B_1 r + \frac{1}{3}B_2 r + \cdots + \frac{1}{3}B_k r \\ &= \frac{1}{3}(B_1 + B_2 + \cdots + B_k)r \\ &\approx \frac{1}{3}\text{S.A.} \cdot r \end{aligned}$$

Now use the volume formula for a sphere (Theorem 10.21).

$$\frac{4}{3}\pi r^3 = \frac{1}{3}\text{S.A.} \cdot r$$

Solving this formula for S.A,

$$\frac{3 \cdot \frac{4}{3}\pi r^3}{r} = \text{S.A.}$$

Simplifying this fraction yields the theorem. ∎

Archimedes derived the formula for the surface area of a sphere by inscribing the sphere inside a circular cylinder with the same diameter and height as the sphere. He was so excited about this result that he directed the sphere in a cylinder diagram to be engraved on his tombstone. He noted (and we note) the following about a sphere with diameter d, which summarizes the formulas for both the circle and sphere.

Great circle: circumference πd, area $\frac{1}{4}\pi d^2$

Sphere: surface area πd^2, volume $\frac{1}{6}\pi d^3$

Notice that the coefficients of area and volume are proportional to the corresponding coefficients ($\frac{1}{2}$ and $\frac{1}{3}$) in formulas for areas of triangles and volumes of pyramids. These results suggest that it might be easier to have students learn the formulas for circles and spheres in terms of their diameters rather than in terms of their radii.

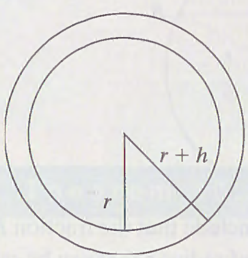
Relationships among formulas

Many calculus students notice that when $V = \frac{4}{3}\pi r^3$, then $\frac{dV}{dr} = 4\pi r^2 = \text{S.A.}$ That is, the derivative of the volume with respect to the radius seems to be the surface area. Also, when $A = \pi r^2$, then $\frac{dA}{dr} = 2\pi r = C$. Are these relationships calculational coincidences or are they the result of special properties of the circle and sphere?

To answer this question, recall that if g is a function of r, then

$$g'(r) = \lim_{h \to 0} \frac{g(r+h) - g(r)}{h}.$$

Figure 73



Let $A(r)$ be the area of a circle of radius r. Then

$$A(r+h) - A(r) = \pi(r+h)^2 - \pi r^2 = 2\pi rh + \pi h^2.$$

The quantity $A(r+h) - A(r)$ is the area of the ring-shaped region between two concentric circles of radii r and $r + h$ (Figure 73). As h approaches 0, it follows algebraically that

$$A'(r) = \lim_{h \to 0} \frac{A(r+h) - A(r)}{h} = \lim_{h \to 0}(2\pi r + \pi h) = 2\pi r = C(r).$$

Geometrically, as the width of the ring gets smaller, the relative change in area of the circle from r to $r + h$ approaches the circumference of the circle of radius r.

The equation $\frac{dV(r)}{dr} = 4\pi r^2 = S(r)$ can be interpreted in a similar way. Let $V(r)$ be the volume of a sphere of radius r. The quantity $V(r+h) - V(r) = \frac{4}{3}\pi(r+h)^3 - \frac{4}{3}\pi r^3$ is the volume of the spherical shell between the spheres of radii r and $r + h$. (A spherical shell is the 3-dimensional counterpart of a 2-dimensional ring.) This quantity can be simplified to

$$V(r+h) - V(r) = 4\pi r^2 h + 4\pi r h^2 + \frac{4}{3}\pi h^3.$$

The first term of this sum, $4\pi r^2 h$, is the approximation to the volume of this shell obtained by multiplying the area $4\pi r^2$ of the inner surface of the shell by the shell's thickness (an underestimate), while the term $4\pi rh^2 + \frac{4}{3}\pi h^3$ is the error resulting from this underestimate of the shell volume.

As h approaches 0, it follows algebraically that

$$V'(r) = \lim_{h \to 0} \frac{V(r+h) - V(r)}{h} = \lim_{h \to 0}\left(4\pi r^2 + 4\pi rh + \frac{4}{3}\pi h^2\right) = 4\pi r^2 = S(r).$$

Geometrically, as the thickness of the shell decreases to zero, the relative change in volume of the sphere between r and $r + h$ approaches the surface area of the sphere of radius r.

So these relationships are not coincidences, but explainable both geometrically and algebraically.

10.3.1 Problems

1. Interpret the formula L.A. $= \pi\ell^2 \sin\phi$ for the lateral area of a right circular cone when $\phi = 90°$.

2. Find a formula for the lateral area of a right circular cone in terms of the radius of its base and the angle between a lateral edge and the plane of its base.

3. a. Find the volume of a right circular cone in terms of the measure of the opening of the cone and its slant height.

b. Use this formula to answer the following question. A sector is cut out of a disk, and the radii that are edges of the sector are made to coincide. Then the disk has become the lateral surface of a right circular cone. What is the measure of the central angle for which the cone's volume is maximized?

4. A rectangle with dimensions a and b can be rolled up in two ways to become the surface area of a cylinder.

a. What are the volumes of the two cylinders so formed?

b. Which way should an 8.5″ by 11″ sheet of paper be rolled up to obtain the cylinder of larger volume?

5. a. Suppose that T_r is an equilateral triangle and S_r is a square, both inscribed in a circle of radius r. Find formulas for the areas $A(r)$ and perimeters $P(r)$ of T_r and S_r in terms of r and show that it is not true that $A'(r) = P(r)$ for either of these figures.

b. Suppose that, in part **a**, T_r and S_r were circumscribed about a circle of radius r. Is it true that $A'(r) = P(r)$ for either figure?

6. Let $A_n(r)$ and $P_n(r)$ be the area and perimeter of a regular n-sided polygon inscribed in a circle of radius r. Investigate the behavior of the derivative $A'_n(r)$ as n increases.

7. Prove: If a cylinder's base is the disk of a great circle of a sphere, and the cylinder's height is equal to the diameter of the sphere, then the cylinder's total surface area is $\frac{3}{2}$ the surface area of the sphere.

8. Show that the surface area $S_{h,r}$ of a spherical cap of depth h on a sphere of radius r is given by

$$S_{h,r} = 2\pi rh.$$

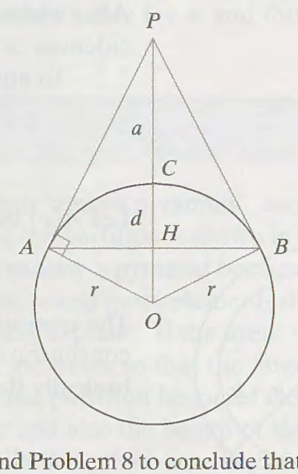
(*Hint*: Use the result of Problem 13 in Section 10.2.3 and an argument similar to that in the proof of Theorem 10.22.)

9. a. Figure 74 displays a cross-section of a sphere with center O by a plane that contains point P and the arc $\widehat{ACB}$ that is visible from P. Let r be the radius of the sphere, let a be the distance from P to the sphere, let H be the intersection of $\overline{AB}$ and $\overline{OP}$, and let $d = HC$. Explain why

$$OH = \frac{r^2}{r + a},$$

and use this fact to show that $d = \frac{ra}{r+a}$.

Figure 74



b. Use part **a** and Problem 8 to conclude that the fraction $F_{a,r}$ of the surface area of a sphere of radius r that can be seen from a point P that is a units above the surface of the sphere is given by

$$F_{a,r} = \frac{a}{2(r + a)}.$$

c. Assume that Earth is a sphere of radius 3960 miles. What fraction of Earth's surface can be viewed from a satellite in a circular orbit 1000 miles above the surface?

d. Explain why 3 satellites placed in geosynchronous circular orbits 20,000 miles above the equator are insufficient to operate a communications network for the entire Earth.

e. Based on the meaning of $F_{a,r}$, how do you expect $F_{a,r}$ to vary as a increases without bound for fixed r? Confirm your expectations mathematically using the formula for $F_{a,r}$ from part **b**.

10.3.2 The Isoperimetric Inequalities

Theorem 10.2 in Section 10.1.1 states a well-known result: Of all rectangles with a given perimeter, the square has the greatest area. The problem this theorem solves is an *isoperimetric problem*, that is, a max-min problem dealing with figures that have the same perimeter. That particular theorem relates to a situation where you have a fixed amount of fence and you wish to enclose the rectangular region with the largest area.

Dido's problem

If we do not restrict the shape of the region, then we are faced with the general problem for the plane. The problem of enclosing the largest possible area with a given perimeter is known as *Dido's problem.* The story of this problem is interesting.

Carthage was a port on the Mediterranean Sea in Northern Africa and one of the great cities of ancient times. A major foe of Rome, Carthage engaged Rome in three major wars in the 3rd and 2nd centuries B.C., being destroyed in the last of these. Carthage later was rebuilt and was important in the Roman empire. It was again destroyed around 450 A.D. and was depopulated after 698 A.D. The ancient site of Carthage is now in the suburbs of the city of Tunis in Tunisia. Archaeological excavations indicate that Carthage was founded around 750 B.C.

The story of Dido as told in the epic poem *Aeneid* by the Roman poet Virgil clearly involves many myths. According to legend (but possibly with some reality), Dido was the daughter of King Belus of the Phoenician city of Tyre. She fled to Africa with some devoted followers after her husband was murdered. She was offered only as much land as she could surround with a bull's hide. Determining the most efficient shape became Dido's problem. Her solution was to cut the hide into very thin strips and lay them out end to end to enclose the largest possible area. The enclosed region became the site of Carthage.

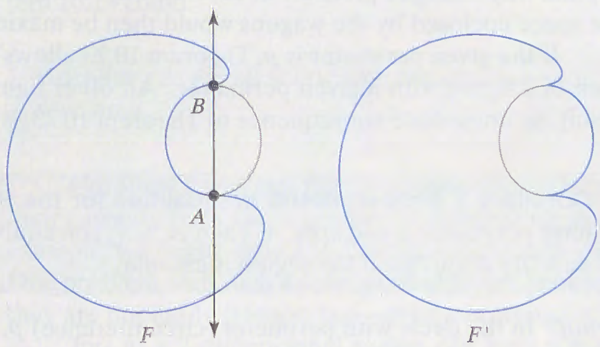
The next theorem states the solution to Dido's problem. A rigorous solution requires a very careful definition of the length of a *simple closed curve*, which we do not give. So we can present only a partial proof. This proof was first given by the Swiss mathematician Jakob Steiner (1796–1863). We assume that there exists a simple closed curve with the largest area for a given perimeter.

Theorem 10.23 Of all plane figures with the same perimeter, the circle has the largest area.

Proof: Let F be the simple closed curve with largest area for a given perimeter p.

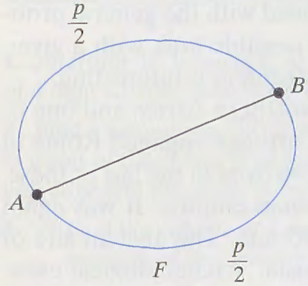
1. First we show by an indirect proof that F must be convex. Suppose F is not convex. Then there exist two points A and B on F such that $\overline{AB}$ contains only points in the exterior of F (Figure 75). Then reflect the nonconvex part of F bounded by A and B over $\overleftrightarrow{AB}$. The new curve F' has the same perimeter as F and greater area. This contradicts F having the largest area for its perimeter.

Figure 75



2. Now we show by indirect proof that there exist lines through any point of F that split the area of F into two equal halves. Suppose F cannot be split in this way. Let A be any point on F. Since F has perimeter p, there is a unique second point B on F whose distance from A along F is $\frac{p}{2}$. (B can be said to be

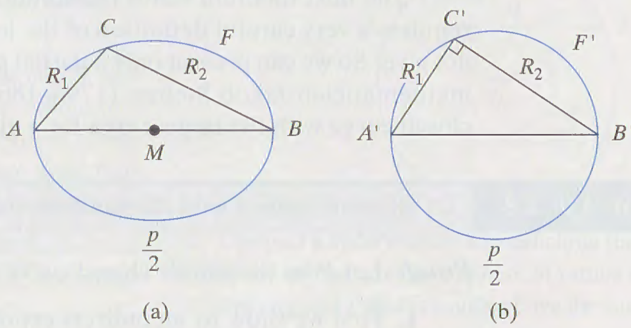
Figure 76



"halfway around" F from A.) Since from part (1) F is convex, $\overline{AB}$ splits F into two regions, each with the same perimeter (Figure 76). If these regions do not have the same area, then we can reflect the region with the larger area over $\overleftrightarrow{AB}$ and replace F by the new curve. But this means there is a curve with perimeter p enclosing a larger area than F, contrary to our assumption. So $\overline{AB}$ splits F into two regions with the same area.

3. Lastly, we show that each half of this split must be a semicircle. Let C be a point on F other than A or B. The half of the region containing C and bounded by $\overline{AB}$ consists of two regions R_1 and R_2 with areas A_1 and A_2 (Figure 77a) and ΔABC with area A_3. If $\angle ACB$ is not a right angle, then consider the new curve F' obtained by attaching regions R_1 and R_2 upon $\Delta A'C'B'$, where $A'C' = AC$, $B'C' = BC$, and $\angle C'$ is a right angle. Since $\alpha(\Delta ABC) = \frac{1}{2}AC \cdot BC \cdot \sin \angle ACB$, and $\alpha(\Delta A'B'C') = \frac{1}{2} \cdot AC \cdot BC$, the area of F' is greater than the area of F. This again contradicts F being the curve with maximal area for its perimeter, so $\angle ACB$ must be a right angle. Let M be the midpoint of $\overline{AB}$. Since ΔACB is a right triangle, M is equidistant from A, B, and C. So C lies on the circle with center M and passing through A and B. Thus all points of F on this side of $\overline{AB}$ are on this circle, and the same argument can be applied to the part of F on the other side of $\overline{AB}$.

Figure 77



This theorem may explain why tepees, hogans, igloos, yurts, and other structures of peoples throughout the world are circular in shape. These structures maximize the floor space that can be surrounded by a fixed amount of material. Since the materials that are used in these structures were valuable and sometimes scarce resources, people learned not to use more material than necessary. The same principle may explain why besieged pioneers in covered wagons would want to "circle the wagons"; the space enclosed by the wagons would then be maximized.

If the given perimeter is p, Theorem 10.23 allows us to calculate the maximum area of a figure with a given perimeter. All other figures must have less area. The result, an immediate consequence of Theorem 10.23, is an *isoperimetric inequality*.

Corollary 1 (Isoperimetric Inequalities for the Plane): Let a plane region have perimeter p and area A. Then $A \leq \frac{p^2}{4\pi}$, or, equivalently, $p \geq 2\sqrt{\pi A}$, with equality occurring if the region is circular.

Proof: In the circle with perimeter (circumference) p, $p = 2\pi r$. Thus $r = \frac{p}{2\pi}$. Since $A = \pi r^2$, substituting for r, we have $A = \pi\left(\frac{p}{2\pi}\right)^2 = \frac{p^2}{4\pi}$. By Theorem 10.23, this is the maximum area figures with perimeter p can have. So for any other figure with area A and perimeter p, $A \leq \frac{p^2}{4\pi}$. Solving for p yields the second inequality in the statement of the corollary.

The second inequality in the corollary provides a useful restatement of Theorem 10.23.

Corollary 2: Of all plane figures with the same area, the circle has the least perimeter.

It is not always desirable to have the smallest perimeter for a given area. Upscale suburban developments outside many cities include human-made lakes. Shoreline property is worth more per square foot than other property, so the developer may try to make the lake longer and thinner. Then the region covered by water will have a larger perimeter than a circle of equal area.

The 3-dimensional counterpart

Many igloos and yurts not only have circular bases but also have roughly hemispherical roofs. These roofs suggest that 3-dimensional counterparts to the 2-dimensional isoperimetric inequalities would involve hemispheres or spheres. In fact, the sphere plays the same role in these inequalities in three dimensions that the circle plays in two dimensions. However, the proof we have given for Theorem 10.23 does not generalize to three dimensions. A proof for the 3-dimensional counterpart was first found by Hermann Amandus Schwarz (1843–1921). We omit the proof.

Theorem 10.24 Of all solids with the same surface area, the sphere has the largest volume.

Theorem 10.24 has corollaries corresponding to those of Theorem 10.23.

Corollary 1 (Isoperimetric Inequalities for Space): Let a solid have surface area A and volume V. Then $V \leq \sqrt{\frac{A^3}{36\pi}}$, or, equivalently, $A \geq \sqrt[3]{36\pi V^2}$, with equality occurring if the region is spherical.

Proof: The proof follows the idea of the proof of Corollary 1 to Theorem 10.23 and is left to you. ∎

Theorem 10.24 can be restated in the language of minimums just as Theorem 10.23 could.

Corollary 2: Of all solids with the same volume, the sphere has the least surface area.

Corollary 2 suggests that containers should be spherical if they are to maximize their capacity for a given amount of material needed to form the container. So, for example, we might expect to have spherical milk cartons or spherical cereal boxes. One problem with such a solution is obvious: Spheres roll! A second problem is that they are not easily transported—even a bowling ball has holes!

Just as a 2-dimensional region can have a large perimeter and small area, a 3-dimensional solid can have a large surface area and small volume. Filters such as those found in heaters and air conditioners, water purifiers, and cigarettes are based on the idea that small particles can stick to surfaces. They have very large surface areas for their volumes.

10.3.2 Problems

1. An equilateral triangle, a square, and a regular hexagon each have perimeter p.

a. Give the area of each figure. How much larger in area is a circle with a circumference of p?

b. A lake has a surface area of 1000 km^2. What are the maximum and minimum lengths of beach this lake might have?

2. An equilateral triangle, a square, a regular hexagon, and a circle each have area A. Find the perimeter of each and show that the circle's perimeter is smallest.

3. Let the sides of a triangle be a, b, and c with c and $a + b$ constant. Under these conditions, prove that the triangle with maximal area is isosceles.

4. A boxing "ring" is actually square-shaped. Design a plan for 2000 people to have seats to view the ring that would allow maximum visibility for the greatest number of people.

5. Suppose a local zoning law requires a floor space of at least 600 square feet of living space for a summer cottage. How should the cottage be shaped to minimize the cost of materials for the walls?

6. A rancher has 300 yards of fencing. He wants to use the fencing to construct a rectangular pen and to divide the pen in two halves to separate the bulls from the cows.

a. Find the dimensions of the pen that provides the largest area for bulls and cows.

b. Show that neither splitting a big square pen in two congruent halves, nor using a pen with two square halves, yields the pen with the largest area.

7. A rectangular dog pen is to be built alongside a house. The house will form one side of the pen; fencing will form the other three sides. What shape pen gives the dog the most play area?

8. Prove Corollary 1 to Theorem 10.24.

9. When a spherical soap bubble lands on a flat surface so that the surface becomes one of the sides of the space enclosed by the bubble, the bubble will assume the shape of a hemisphere. (There is soap film on the surface when this happens.) If the radius of the original bubble was r, what is the radius of this hemisphere?

*10. Containers for many foods are circular cylinders. It is in the interest of food distributors to use cylinders with shapes that maximize volume for their surface areas.

a. What ratio of height to radius maximizes the volume of a circular cylinder with a given surface area S?

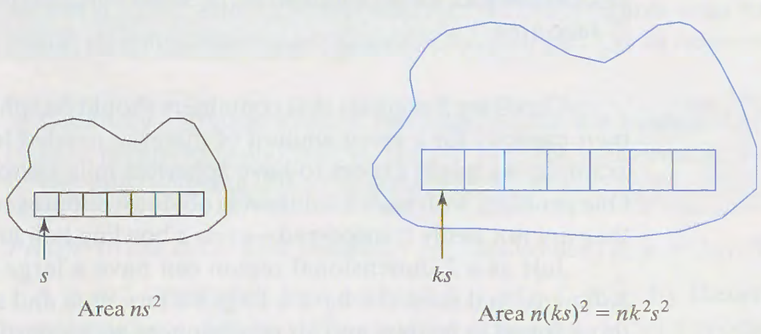
b. As the ratio of height to radius changes, by how much does the volume of a circular cylinder with a given surface area S change?

10.3.3 The Fundamental Theorem of Similarity

From the basic properties of similarity transformations, when two figures are similar with ratio of similitude k, corresponding angles have the same measure and corresponding distances are in the ratio k. But what happens to area and volume?

Consider a square with side s. Under a similarity transformation with ratio of similitude k, the image square has side ks. The original square, which had area s^2, gives rise to an image with area $(ks)^2$, or k^2s^2. The area of a figure that is not a square is determined by partitioning that figure into squares or parts of squares. If each square in the partition of a preimage figure has its area multiplied by k^2 in the image figure, then the area of the image is k^2 times the area of the preimage. Thus if two figures are similar with ratio of similitude k, then the ratio of their areas is k^2 (Figure 78).

Figure 78



In 3-dimensional space, under a similarity transformation with ratio of similitude k, a cube with edge e gives rise to an image cube with edge ke. The original cube, which had volume e^3, gives rise to an image with volume $(ke)^3$, or k^3e^3. Using the same reasoning as in the 2-dimensional situation, if two figures are similar with ratio of similitude k, then the ratio of their volumes is k^3.

These properties are summarized in a theorem we call the Fundamental Theorem of Similarity because of its wide-ranging scope and applications.

Theorem 10.25 **(Fundamental Theorem of Similarity):** If β is the image of α under a similarity transformation of magnitude k, then

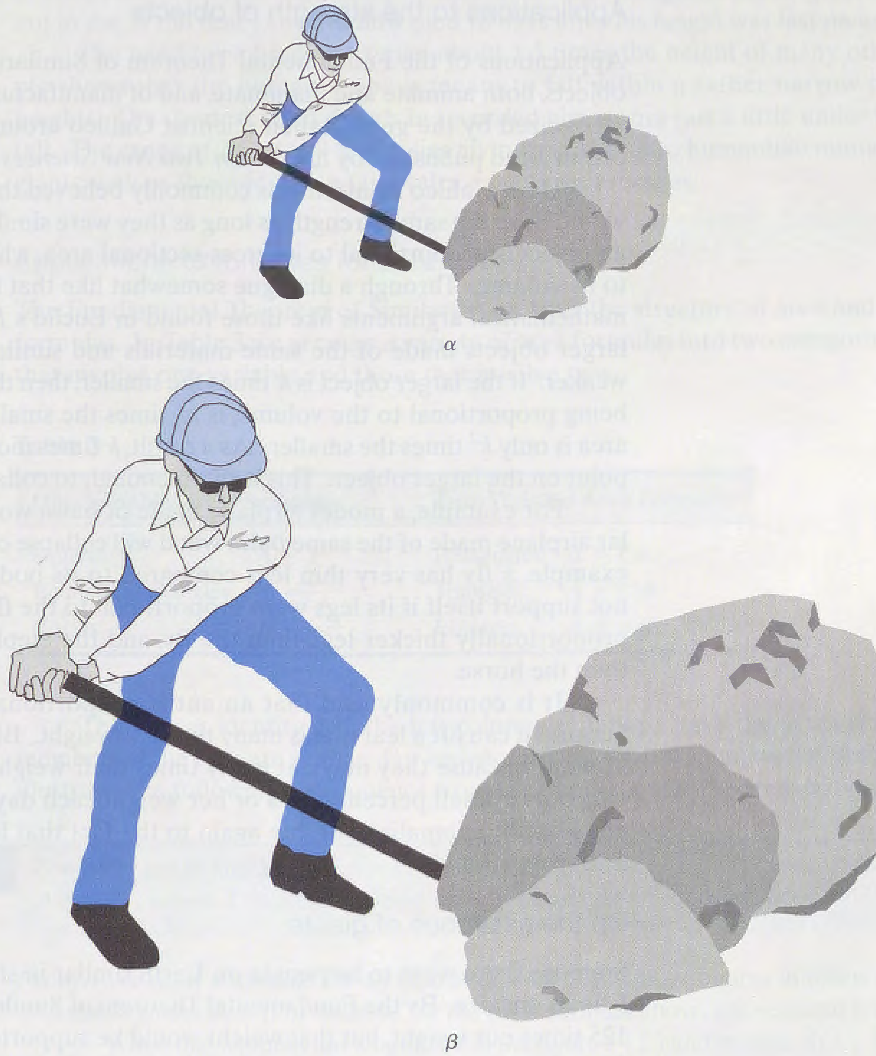
a. angle measures in β are equal to corresponding angle measures in α;

b. distances in β are k times corresponding distances in α;

c. areas in β are k^2 times corresponding areas in α; and

d. volumes in β are k^3 times corresponding volumes in α.

Figure 79



For instance, the ratio of similitude in Figure 79 is 1.75. So, while corresponding angle measures are equal, the lever in β is 1.75 times as long as the lever in α, and β occupies an area 1.75^2 times as great on the page as α does.

The parts of the Fundamental Theorem of Similarity are instances of an important pattern. Rewrite part (b) as *distances in* β *are* k^1 *times corresponding distances in* α. Then the exponent of k in parts (b), (c), and (d) signifies the dimension of the measure. Distance is 1-dimensional, area is 2-dimensional, and volume is 3-dimensional. This pattern also extends to angle measure. Part (a) can be written as *angle measures in* β *are* k^0 *times corresponding angle measures in* α. This suggests that angle measure is 0-dimensional, a property that is corroborated by examination of formulas that involve angle measures. One such formula is $s = r\theta$ for the length s of an arc with a central angle of θ radians in a circle with radius r. If θ had any dimension, then the lengths r and s would have different dimensions.

Another argument for angle measure being 0-dimensional is that angle measure is defined around a point, a 0-dimensional figure. From this perspective, the other dimensions fit in familiar ways. Distance and length are defined along a line, a 1-dimensional figure; area is defined on plane figures, which are 2-dimensional; and volume deals with the capacity of a figure in 3-dimensional space.

Applications to the strength of objects

Applications of the Fundamental Theorem of Similarity to the strength of natural objects, both animate and inanimate, and of manufactured objects as well, were first recognized by the great Italian scientist Galileo around the beginning of the 17th century and published by him in *On Two New Sciences* in 1638.

Until Galileo's time, it was commonly believed that larger and smaller objects would have the same strength as long as they were similar. However, the strength of an object is proportional to its cross-sectional area, while its weight is proportional to its volume. Through a dialogue somewhat like that found in Plato's writings and mathematical arguments like those found in Euclid's *Elements*, Galileo shows why larger objects made of the same materials and similar to smaller objects will be weaker. If the larger object is k times the smaller, then the weight of the larger object, being proportional to the volume, is k^3 times the smaller, while the cross-sectional area is only k^2 times the smaller. As a result, k times more pressure is placed on each point on the larger object. This is often enough to collapse the larger object.

For example, a model airplane made of balsa wood can fly, but a larger similar airplane made of the same balsa wood will collapse of its own weight. As another example, a fly has very thin legs compared to its body, but a larger animal could not support itself if its legs were proportional to the fly's. So the horse must have proportionally thicker legs than the fly, and the elephant must have thicker legs than the horse.

It is commonly said that an ant is proportionally stronger than a human because it can lift a leaf that is many times its weight. Birds are said to eat more than humans because they may eat many times their weight in a day, whereas a human eats only a small percent of his or her weight each day. These biological phenomena are not anomalies but due again to the fact that length, area, and volume are not proportional.

On the existence of giants

Suppose there were to be people on Earth similar in shape and substance to us but 5 times our size. By the Fundamental Theorem of Similarity, such people would have 125 times our weight, but that weight would be supported on only 25 times the area. As a result, each part of the body would have to support 5 times as much weight as our parts do. Even champion weightlifters seldom lift more than twice their body weight, and when they do, it is only for a few seconds. The bodies of such giants would collapse under their own weight.

Reality supports the theory. You may have seen television or movie pictures of people weighing well over 500 pounds. Even though their legs are wider, these people often can barely move. They cannot support their own weight.

Very heavy people tend not to have the same shape as lighter people. So they do not necessarily tell us how tall people can become without changing shape. Experience suggests that the answer is not much more than 1.5 times average height. The tallest man on record from anywhere in the world has been Robert Wadlow (1918–1940). Wadlow was born in Alton, Illinois, and was known from his youth as a giant, and so his condition was subject to medical study and there exist quite a number of pictures of him. He was growing his entire life, and on June 27, 1940, he was measured to have a height of 8 feet, 11.1 inches.

If you saw a picture of Wadlow without other people next to him, you would have little idea that he was not of average height. But he was so large that he could not support his weight without help and needed a leg brace for support. This tragically turned out to be the cause of his death. While getting out of a car—a difficult task for a man of such size—his brace cut a deep wound in his leg. Gangrene set in (penicillin was not in use at the time) and Wadlow died 18 days after his height was last measured.

The need for a brace by a man about 1.5 times the height of many other people shows that the human shape is meant to fall within a rather narrow range of heights. The shortest adult dwarfs in recorded history are just a little under two feet tall. The range of 2 feet to 9 feet belies giving any reality to humanlike miniatures or giants such as those found in fairy tales, cartoons, or movies.

Applications to formulas for area

The Fundamental Theorem of Similarity explains the structure of area and volume formulas. In Table 3 we arrange a variety of area formulas into two categories: those that involve one variable and those that involve two.

Table 3

One-Variable Area Formulas		Two-Variable Area Formulas	
Squares:	$A = s^2$	Rectangles:	$A = \ell w$
Equilateral triangles:	$A = \frac{s^2\sqrt{3}}{4}$	Triangles:	$A = \frac{1}{2}bh$
Circles:	$A = \pi r^2$	Ellipses:	$A = \pi ab$

The figures identified in the left column of Table 3 have the property that all members of the type are similar. For instance, all equilateral triangles are similar. This illustrates the following consequence from the Fundamental Theorem of Similarity.

Theorem 10.26 For every set of similar 2-dimensional figures, there is an area formula of the form $A = kL^2$, where L is a corresponding length on one of the figures.

Proof: We want a formula for all figures of a set S of similar figures in terms of a particular segment in S. (For instance, for the circle formula above, this segment is a radius.)

When this segment has length L, call the figure $F(L)$ and its area $A(L)$. Consider the figure $F(1)$. This is the figure in which $L = 1$. (For the circle, this is the unit circle.)

Every figure $F(L)$ in the set is the image of $F(1)$ under a similarity transformation with a magnitude L. So, by the Fundamental Theorem of Similarity, $A(L) = A(1)L^2$. ┘

In practice, L is usually a natural length on a figure, such as a diameter, height, or side. The proof of Theorem 10.26 does more than prove the statement of the theorem. It shows that the constant k in the statement of the theorem is the area of the figure in the set when $L = 1$. For instance, the constant π in the expression πr^2 is the area of a circle when its radius is 1. The constant $\frac{\sqrt{3}}{4}$ in the formula for the area of an equilateral triangle is the area of an equilateral triangle with side $s = 1$. Thus, the formula for a set of similar figures in terms of a particular segment on those figures is entirely determined by the area when that segment has length 1.

Applications to formulas for volume

Volume formulas are slightly more complex, for they may involve one, two, or three variables. But the idea of Theorem 10.26 still holds. Examine the volume formulas shown in Table 4.

Table 4

One-Variable Volume Formulas	Two-Variable Volume Formulas	Three-Variable Volume Formulas
Cubes: $V = s^3$	Circular cylinders: $V = \pi r^2 h$	Boxes: $V = \ell wh$
Spheres: $V = \frac{4}{3}\pi r^3$	Circular cones: $V = \frac{1}{3}\pi r^2 h$	

Theorem 10.27 For every set of similar 3-dimensional figures, there is a volume formula of the form $V = kL^3$, where L is a corresponding length on one of the figures.

We leave the proof of Theorem 10.27 to you. We also leave it to you to describe the quality of figures that separates those whose volume formulas involve two variables from those whose volume formulas require three variables.

10.3.3 Problems

1. a. A figurine weighs w kg. A similar figurine, made of the same materials, is twice the height of the first figurine. What is the weight of the larger figurine?
   b. Suppose a pail filled with sand weighs 5 lb. A similar pail, twice the height, is filled with sand. What is its weight?
   c. Generalize parts **a** and **b**.
2. According to 1998 weight guidelines from the National Institutes of Health, the middle weight of the healthy range for a 6-foot person of either sex is 158 lb.
   Use the ideas of this section to determine the corresponding weight for a similar person with the given height.
   a. 7 feet b. 5 feet c. h feet
3. Find the area of a circle with diameter 1. Use this result to find a formula for the area of a circle in terms of its diameter.
4. Find a formula for the area of a regular hexagon.
5. Find a formula for the area of a regular n-gon in terms of its side length s. (You will need trigonometry.)
6. In *On Two New Sciences*, Galileo proves the following theorem: The area of a circle is the geometric mean of the areas of two regular n-gons with the same number of sides, one circumscribed about the circle and one with the same perimeter as the circle.
   a. Prove this theorem for a particular value of n.
   b. Prove the general theorem.
7. Does Theorem 10.26 apply also to formulas for surface area? Explain why or why not.
8. Prove Theorem 10.27.
9. Find a formula for the volume of a sphere with diameter d.
10. Find a formula for the volume of a regular square pyramid with base of side s and height h.
11. Find a formula for the volume of a cone with height h and an elliptical base having axes of lengths a and b.
12. Give some volume formulas that involve two variables, and other volume formulas that involve three variables. What quality separates the figures in one group from the figures in the other?

10.3.4 Fractional dimension

Normally we think of dimension as a count of independent directions. For instance, if a figure extends in two directions, the figure is 2-dimensional. A direction such as "13° north of west" is not independent of "north" and "west" because we can go a certain distance north and a certain distance west to get to 13° north of west. But going north and west will never get us in an up or down direction, and so we view up or down as adding a 3rd dimension. If dimension is conceived as a count of directions, there can never be a fractional dimension.

However, recall that the Fundamental Theorem of Similarity (Theorem 10.25) has an interpretation closely linked to dimension. Its four parts can be combined into one statement: *If β is the image of α under a similarity transformation with magnitude k, then d-dimensional measures on β are k^d times corresponding d-dimensional measures on α.* In this form, the dimension d has a role as an exponent. Because exponents do not have to be integers, this form allows for the possibility of fractional dimension.

Another way to connect measures with dimension is found in unit conversion. For instance, we might begin with the conversion of feet to yards.

$$\begin{aligned} 1 \text{ yard} &= 3 \text{ feet} \\ 1 \text{ square yard} &= 9 \text{ square feet} \\ 1 \text{ cubic yard} &= 27 \text{ cubic feet} \end{aligned}$$

Let us rephrase these conversions using exponents.

$$\begin{aligned} 1 \text{ yard} &= 3^1 \text{ feet} \\ 1 \text{ square yard} &= 3^2 \text{ square feet} \\ 1 \text{ cubic yard} &= 3^3 \text{ cubic feet} \end{aligned}$$

Again the dimension of the unit is an exponent, in this case, an exponent in the conversion factor.

Might there be a kind of yard that does not convert to feet by an integer power of 3? The answer is that there is, and for this we consider a real problem.

The length of a coastline

Surveys measure lengths of coastlines. For instance, the National Oceanic and Atmospheric Administration (NOAA) of the U.S. Department of Commerce provides a length of coastline for every state bordering on one of the oceans. The four states with the most coastline are: Alaska, 5580 miles; Florida, 1350 miles; California, 840 miles; Hawaii, 750 miles. It is not easy to calculate the length of a coastline. How far in should you go inland, if at all, on a river? When should an inlet be counted as coastline? Should a small finger of land be included?

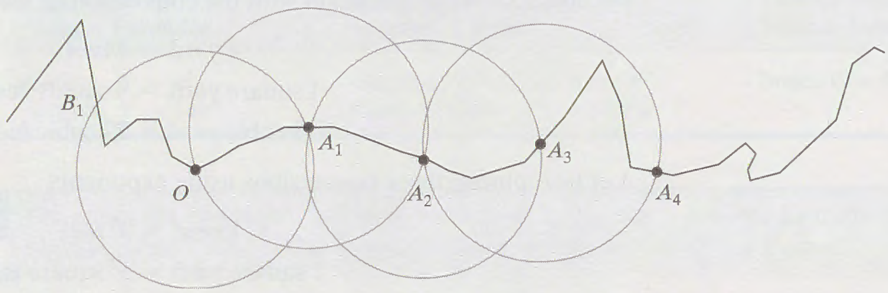
The NOAA figures were calculated in the following way: The general outline of the seacoast as found on charts as near the scale of 1:1,200,000 as possible was used. One inch on such a map is about 19 miles. Measurements were made with a unit measure of 30 minutes of latitude. Coastline of sounds and bays were included to a point where they narrow to a width of 30 minutes of latitude, and then the distance across at this point was used.

A key aspect of this description is that a unit of measure is selected. The unit of measure is critical because the smaller the unit, the more the measurer has to go into the narrower parts of sounds and bays. Fingers of land that might not be considered with a larger unit must be included with a smaller unit. Consequently, *when a smaller unit is used, the coastline is longer*.

This goes against the intuition that a coastline should have a definite length. But it agrees with another intuition, that measuring a coastline is difficult because a coastline really does not have a length at all. In between is a mathematical way of dealing with both intuitions: *coastlines have a measure but it is not 1-dimensional.*

The NOAA figures were calculated from maps, but suppose we were to actually go out to the coast to measure the coastline. Then we might proceed as follows. First we choose a unit, say 1500 meters. We put down a stake at a point O on the coast and draw an imaginary circle with O at the center and a radius of 1500 m. Assuming we are not on a small island, the circle intersects the coast at two points, one on each side of O. Let us call these points A_1 and B_1 (Figure 80). Now we go to A_1, put down a stake, and draw a second circle, this time with A_1 as center and with radius 1500 m. This circle intersects the coastline at O and a new point A_2. We continue this process with each new A_i as center intersecting the coast at A_{i-1} and a point A_{i+1}, until we have gone all the way around (if we are on a large island) or we reach the end of the territory whose coastline concerns us. If O were not at the edge of a territory, we would also need to use point B_1 and generate points B_i in the direction of B_1 just as we generated the points A_i.

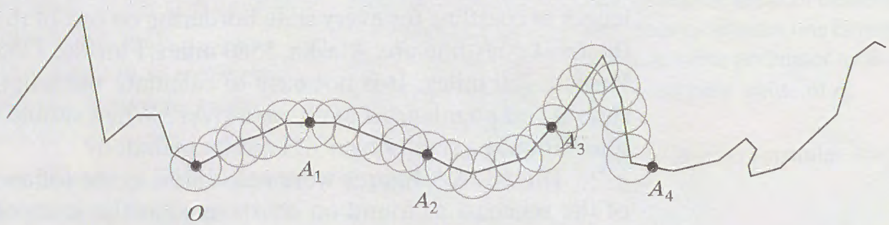
Figure 80



When we have finished covering the coastline, we count the number of spaces between stakes we have put down, and we arrive at a length of the coastline *for the unit of 1500 meters.*

But suppose we chose a unit of 300 meters. If we were measuring the length of a straight road, we would get 5 times the number of 300-meter intervals as we got for the 1500-meter interval. But a coastline is not a road. It goes in and out, and using 300-meter intervals we go in and out more than we did with the 1500-meter intervals. It would not take much of a disruption in the coastline for there to be six 300-meter intervals for each 1500-meter interval, on the average (Figure 81).

Figure 81



In such a case, if the coastline was 210 units long for the unit of 1500 meters, then it would be 1260 units long for each unit of 300 meters. With the unit 1500 meters long, the coastline would be $210 \cdot 1500$, or 315,000 meters long. But with the unit 300 meters long, the coastline would be $1260 \cdot 300$, or 378,000 meters long. Instead of

$$5 \text{ 300-meter units} = 1 \text{ 1500-meter unit}$$

we have

$$6 \text{ 300-meter coastline units} = 1 \text{ 1500-meter coastline unit.}$$

The dimension d of the coastline units is found by solving $5^d = 6$. The solution to this equation is $\log_5 6$, or $\frac{\log 6}{\log 5}$, or about 1.113. This hypothetical coastline would have dimension about 1.113.

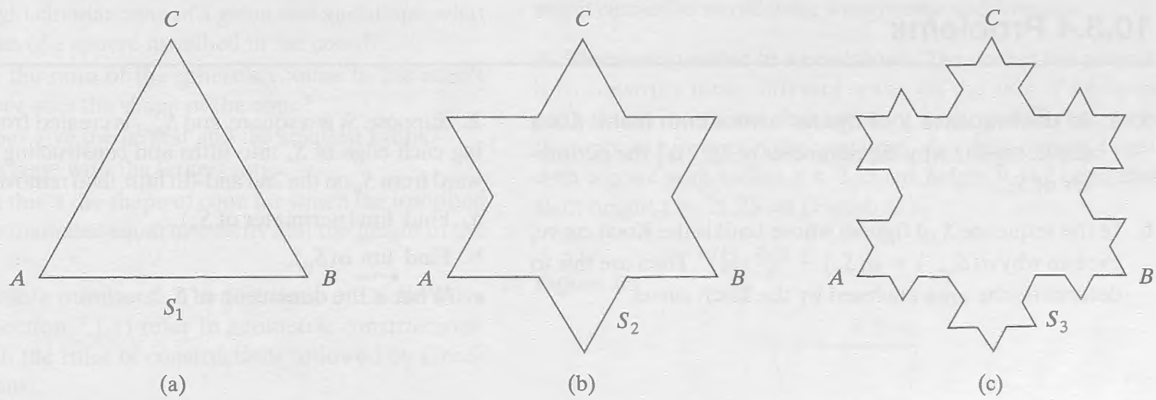
The Koch curve

Study of the *Koch curve* enables us to connect the ideas of similarity, change of units, and dimension. The **Koch curve** is the limiting boundary of a sequence S_n of plane figures created by the following iterative procedure.

S_1 is an equilateral triangle (Figure 82a).

S_{n+1} is created from S_n in the following way. Divide each side of S_n into thirds. On the middle third, construct an equilateral triangle outward from the existing figure, then remove the middle third. The nonconvex polygon S_{n+1} created in this way has 4 times as many sides as S_n, and each side is $\frac{1}{3}$ the length of S_n. Figure 82b pictures S_2; Figure 82c pictures S_3.

Figure 82



Viewed as a 1-dimensional object, the Koch curve has an interesting property. If S_1 has sides of length 1, then the perimeter of S_1 is 3. From the way that S_{n+1} is created from S_n, S_{n+1} has 4 times the number of sides as S_n, and each side of S_{n+1} is $\frac{1}{3}$ the length of a side of S_n. So the perimeter of S_{n+1} is $\frac{4}{3}$ the perimeter of S_n. The sequence of perimeters of $S_1, S_2, \ldots$ begins $3, 4, \frac{16}{3}, \frac{64}{9}, \ldots$. It is a geometric sequence with first term 4 and constant ratio $\frac{4}{3}$, and has nth term $3 \cdot \left(\frac{4}{3}\right)^{n-1}$. This sequence grows without bound, indicating that the limit, the Koch curve, has infinite perimeter. On the other hand, the Koch curve encloses a finite area that can be calculated (see Problem 1b).

The seemingly paradoxical situation of a curve enclosing a finite area but having an infinite length can be explained by the following argument that the Koch curve is not a 1-dimensional figure. Think of the Koch curve as a coastline whose dimension is not known. First use a unit of length 1 to measure the curve, starting at any vertex of the equilateral triangle. With this unit, the curve has length 3 because all of the infinitely many ins and outs are missed by such a large unit.

Next use a unit of length $\frac{1}{3}$. This unit picks up the ins and outs of S_1 but none of the ins and outs of any of the later curves. With this unit, the perimeter is 12, so the total length is 4. Continuing this process, each time picking a unit $\frac{1}{3}$ the size of the previous unit, the perimeter is multiplied by 4, instead of being multiplied by 3 as it would if this were a normal 1-dimensional curve. The situation can be thought of as follows:

$$3 \text{ small units of length} = 1 \text{ large unit of length}$$
$$4 \text{ small units of Koch curve} = 1 \text{ large unit of Koch curve.}$$

Thus the dimension of the Koch curve is the solution to the equation $3^d = 4$, from which $d = \frac{\log 4}{\log 3} \approx 1.26$.

The Koch curve, like many curves of fractional dimension, is *self-similar*, meaning that if a small part of the curve is blown up, the result cannot be distinguished from the original. Specifically, if a part of the Koch curve is blown up 3 times, then it looks like the (new) picture is merely covering 3 times as much as the original curve. When the whole curve is blown up 3 times, however, if its boundary is measured as a coastline is, then the boundary is found to be 4 times what it was. Again this tells us that the dimension d satisfies $3^d = 4$.

The Koch curve is an example of a *fractal*, a term coined by Benoit Mandelbrot in 1976. Since Mandelbrot's path-breaking book, a large literature about fractals has appeared. The study of fractals involves ideas from virtually every branch of mathematics: complex numbers, analysis, topology, geometry, random processes. While not always accessible with elementary mathematics, some of the pictures created using fractals are among the most beautiful ever created with mathematics.

10.3.4 Problems

1. a. In the sequence S of figures whose limit is the Koch curve, explain why the perimeter of S_{n+1} is $\frac{4}{3}$ the perimeter of S_n.

   b. In the sequence S of figures whose limit is the Koch curve, explain why $\alpha(S_{n+1}) = \alpha(S_n) + \frac{3\sqrt{3}}{16} \cdot \left(\frac{4}{9}\right)^n$. Then use this to determine the area enclosed by the Koch curve.

2. Suppose S_1 is a square, and S_{n+1} is created from S_n by dividing each edge of S_n into fifths and constructing a square outward from S_n on the 2nd and 4th fifth, then removing these fifths.

   a. Find $\lim_{n\to\infty}$ (perimeter of S_n).

   b. Find $\lim_{n\to\infty} \alpha(S_n)$.

   c. What is the dimension of S_n?

Chapter Projects

1. Areas of regions with parabolic boundaries. Consider a point P inside a parabola, and lines through this point. Each line cuts off a region inside the parabola. Does the midpoint principle (Section 10.1.3) apply to the problem of finding the line that cuts off the smallest area? (Use any method to explore this question.) Answer this question for points inside other curves: circle, ellipse, hyperbola.

2. The method of exhaustion. Investigate Archimedes' development of the method of exhaustion, and demonstrate its connection to integral calculus.

3. Quadrature of the parabola. Archimedes was able to find the area of a region between a segment and a parabola. This *quadrature of the parabola* was one of the developments that later mathematicians utilized in the evolution of calculus. Find a source that explains how Archimedes did this, and rewrite his method in your own words.

4. Area under a cycloid. In 1637, Gilles Persone de Roberval (1602–1675) devised an elegant geometric proof that the area under one arch of the cycloid

(1) $\quad x = a(t - \sin(t)) \qquad y = a(1 - \cos(t))$

is $3\pi a^2$ by applying Cavalieri's Principle twice. Complete the following parts to see how Roberval's proof works.

a. Use a graphing utility to plot the first half of the arch of the cycloid described in (1), the left half of the circle generating that cycloid,

(2) $\quad x(t) = -a\sin(t) \qquad y(t) = a(1 - \cos(t))$,

and the curve with parametric equations

(3) $\quad x(t) = -at \qquad y(t) = a(1 - \cos(t))$,

all for $a = 1$, on the same coordinate axes.

(Roberval referred to the curve (3) as the *companion* of the cycloid (1).)

b. At any given value t in the plotting interval $0 < t < \pi$, explain why the corresponding points P_s on the semicircle, P_c on the companion curve, and P_t on the cycloid are all on the same horizontal line L_t, and that the distance between P_s and the y-axis is equal to the distance between the points P_c and P_t.

c. Use part **b** and Cavalieri's Principle to conclude that the area between the companion curve and the cycloid is $\frac{\pi a^2}{2}$.

d. Consider the rectangle R joining the points $(0,0)$, $(\pi a, 0)$, $(\pi a, 2a)$ and $(0, 2a)$. Prove that for any given t with $0 < t < \pi$, the point P_t is the same distance from the right side of R as the point $P_{\pi - t}$ is from the left side of R.

e. Use part **d** and Cavalieri's Principle to conclude that the companion curve (3) divides the rectangle R into two regions of equal area.

f. From parts **c** and **e**, show that the area under one arch of the cycloid (1) is $3\pi a^2$.

5. Squaring the circle. One of the classic problems of antiquity was "squaring the circle". Geometers sought to construct a square with area equal to that of a given circle, using straightedge and compass. It was not proved until the 19th century that, with those tools, the problem cannot be solved. Investigate the history of this problem, and write an essay explaining why it cannot be solved and then describing how it can be solved, if the restriction of using the tools of antiquity is removed.

6. A sphere inscribed in a cone.

a. Given a right circular cone of a given size and shape, what is the radius of a sphere inscribed in the cone?

b. How does the ratio of the sphere's volume to the cone's volume vary with the shape of the cone?

c. What shapes give the largest and the smallest ratios?

d. Sketch the cone with the largest ratio.

e. Prove that this is the shape of cone for which the inscribed sphere has diameter equal to exactly half the height of the cone.

7. Constructible numbers. The first three of Euclid's five postulates (Section 7.1.1) refer to geometric constructions. They establish the rules of constructions followed by Greek mathematicians.

1. Given two points P and Q in the plane, the line $\overleftrightarrow{PQ}$ can be constructed.
2. Given a point P and a given line segment $\overline{RS}$, a circle can be constructed with center P and radius RS.

The instruments used for these constructions are (1) the straightedge and (2) the compass. A point may be used as a given point in a construction if and only if it is given or is an intersection point of constructed figures, that is, an intersection point of lines and/or circles. (This means that in a classical Greek construction, you cannot merely open a compass to any radius.) A **Euclidean construction** of a geometric figure F is an algorithm that begins with given geometric objects (points, lines, line segments, triangles, circles, etc.) and proceeds in a finite number of allowable compass and straightedge constructions to the figure F. A real number c is **constructible** if and only if a line segment of length $|c|$ can be obtained from a given line segment of length 1 by a Euclidean construction.

a. Given that a and b are constructible numbers, prove that the numbers $a+b$, $a-b$, ab, $\frac{a}{b}$, and $\sqrt{|a|}$ are constructible.

b. Let E be the set of all constructible numbers. Prove that $\langle \mathrm{E}, +, \cdot \rangle$ is a field containing both the subfield $\langle \mathbf{Q}, +, \cdot \rangle$ of rational numbers and the subfield $\langle \mathbf{Q}(\sqrt{\mathbf{a}}), +, \cdot \rangle$ for any constructible number a that is not the square of a rational number.

c. Suppose that P and Q are points in the plane whose coordinates are in E and that r is in E. Prove that $\overleftrightarrow{PQ}$ and the circle with center P and radius r have equations whose coefficients are in E. Call such lines and circles *constructible lines* and *constructible circles*.

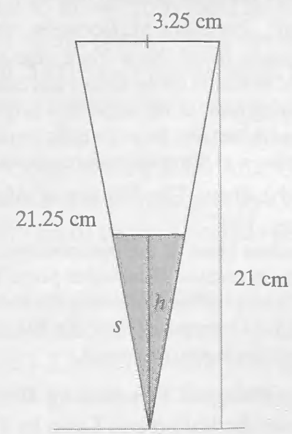
d. Prove that the coordinates of all points of intersection of two constructible lines, or a constructible line and a constructible circle, or two constructible circles are in E.

e. Explain why every constructible number is an algebraic number.

8. Duplicating the cube. Another classic problem of antiquity was "duplicating the cube". Geometers sought to construct a cube with volume equal to twice that of a given cube, using straightedge and compass. It was not proved until the 19th century that with those tools, the problem cannot be solved. Solutions to the problem, without the restriction to the tools of antiquity, were found by Erastosthenes, Menaechmus, and Nicomedes, among others. Investigate the history of this problem, and explain why it cannot be solved using straightedge and compass.

9. Measuring water in a container. The goal of this project is to construct three different scales on the side of an open cone-shaped container that measure in three different ways the amount of water in the container. For definiteness, begin with a cone with radius $r = 3.25$ cm, height $h = 21$ cm, and slant height $s = 21.25$ cm (Figure 83).

Figure 83



a. *Depth scale:* On a strip of paper 21.25 cm long, make a "depth scale" that can be pasted to the slanted side of the cone. The problem is to do this in such a way that a mark s on the scale measures the depth h of the water in the cone, with a mark for every centimeter of depth. Describe the functional relationship between s and h.

b. *Volume scale:* On a second strip of paper 21.25 cm long, make a "volume scale" that can be pasted to the slanted side of the cone. Do this in such a way that a mark s on the scale measures the volume V of the water in the cone, with a mark for every 10 mL of volume. Describe the functional relationship between s and V.

c. *Rainfall scale:* On a third strip of paper 21.25 cm long, make a "rainfall scale" that can be pasted to the slanted side of the cone. The idea here is that during a rainfall of d centimeters the cone will fill to a certain level. Make the scale in such a way that a mark s on the scale measures the amount d of rainfall, with a mark for every centimeter of rainfall. Describe the functional relationship between s and d.

10. The Mandelbrot set. Perhaps the most famous fractal is the Mandelbrot set, the graph of a set of complex numbers.

a. Explain how to determine whether a particular complex number is in the set, and give examples of numbers that are in the set and not in the set.

b. Explain why the Mandelbrot set is an example of a fractal.

c. Describe some of the properties of the Mandelbrot set.

Bibliography

Unit 10.1 References

Bailey, David H., Jonathan M. Borwein, and Peter B. Borwein. "Ramanujan, Modular Equations, and Approximations to Pi, *or* How to Compute One Billion Digits of Pi." *American Mathematical Monthly* 96(3):(1989) 201–219.

Bailey, David H., Jonathan M. Borwein, Peter B. Borwein, and S. Plouffe. "The Quest for Pi." *The Mathematical Intelligencer* 19(1):(1997) 50–57.

These two articles provide insight into the mathematics and the drama in the search to know ever more decimal places of π. The first article is reprinted in the Berggren reference below.

Beckmann, Petr. *A History of Pi.* Boulder, CO: Golem Press, 1971.

A broad and very readable detailed history.

Berggren, Lennart, Jonathan M. Borwein, and Peter B. Borwein. *Pi: A Source Book.* New York: Springer, 1997.

A compilation of 70 articles on the history and calculation and attraction of π, including most of the important original articles on the developments with respect to π, ranging from attempts by the ancients to estimate π to the most modern calculations of digits of π.

Fauvel, John, and J. Gray. *The History of Mathematics.* London: Macmillan, 1987.

This work examines some of the contributions of history's most renowned mathematicians. It includes proof and discussion of many of Euclid's propositions, including the construction of a parallelogram that has area equal to that of a triangle. Quadrature of circles and parabolas is also discussed.

Galilei, Galileo. *Dialogues Concerning Two New Sciences.* Translated from the Italian and Latin by Henry Crew and Alfonson de Salvi. New York: Macmillan, 1933.

This original text of Galileo discusses the relationships between height, weight, volume, and strength of similar-shaped objects.

Moise, Edwin E. *Elementary Geometry from an Advanced Standpoint.* Reading, MA: Addison-Wesley, 1963.

Chapter 13 defines an area function and then deduces well-known area theorems. Chapter 14 shows that the same area function can be deduced when the area theorems are taken as postulates.

Monte Carlo methods:
<http://csep1.phy.ornl.gov/mc/node1.html>

Unit 10.2 References

Altshiller-Court, Nathan. *Modern Pure Solid Geometry.* New York: Macmillan, 1935.

This text provides an in-depth look at properties of the tetrahedron. Of particular interest are several proven results on volume.

Courant, Richard, Herbert Robbins, and Ian Stewart. *What Is Mathematics?* New York: Oxford University Press. New Edition, 1996.

The text contains history and discussion of the problems of squaring the circle and duplicating the cube.

Unit 10.3 References

Emert, John, and Roger Nelson. "Volume and Surface Area for Polyhedra and Polytopes." *Mathematics Magazine* 70, December 1997, 365–371.

This article explores surface area as the derivative of volume with respect to inner radius. It is shown that this result generalizes to a wide range of n-dimensional figures.

Mandelbrot, Benoit B. *The Fractal Geometry of Nature.* San Francisco: W. H. Freeman, 1982.

This is the major work by the mathematician who coined the term "fractal" and helped to popularize the theory of fractals.

Stein, Sherman. *Archimedes: What Did He Do Besides Cry Eureka?* Washington, DC: Mathematical Association of America, 1999.

This book describes in some detail the work of the person many feel was the greatest mathematician of antiquity. Included are his work with the method of exhaustion and the quadrature of the parabola.

Project Reference

Peressini, Anthony L., and Donald R. Sherbert. *Topics in Modern Mathematics for Teachers.* New York: Holt, Rinehart and Winston 1971.

Chapter 9 of this book gives additional information on Euclidean constructions.

Chapter 11

AXIOMATICS AND EUCLIDEAN GEOMETRY

Before the work of Thales, Pythagoras, Euclid, Archimedes, and other mathematicians of the ancient Greek empire, existing records from Babylonia and Egypt show an understanding of *local deduction*, whereby a proposition is logically deduced from other propositions and principles. For instance, the Babylonians had derived a form of what we now call the Quadratic Formula by using mathematical principles not much different from those we would use today.

The Greek mathematicians were the first to move from local deduction to *global deduction*, in which every proposition is part of the same logical system. The global system that these mathematicians developed 2000 years ago is, aside from language, essentially the same system that we use today to study Euclidean geometry. We are only in recent decades beginning to understand the extent of trade and communication among peoples of Europe, Asia, and Africa before 1000 A.D. Geometry done later in Japan, China, and India seems to have been influenced by Greek geometry to such an extent that we can say that virtually all of today's formal geometry worldwide owes its roots to these mathematicians.

Every mathematical system emanates deductively from (1) *undefined terms*, (2) other terms that are defined from the undefined terms, and (3) *axioms*, assumed relationships among these terms. (In Euclid's *Elements*, both postulates and common notions are axioms.) The *theorems* of the system are the propositions that are logically deduced from the axioms.

In 1882 the German mathematician Moritz Pasch (1843–1930) gave the first axiomatic development of Euclidean geometry that would be considered rigorous by today's standards. By the early twentieth century, a multitude of similar, but different, axiom systems were produced for Euclidean as well as other geometries, by such noted mathematicians as David Hilbert (1862–1943), Mario Pieri (1860–1913), Giuseppe Peano (1858–1932), and Oswald Veblen (1880–1960), to name a few. The creation of new axiom systems for geometry continues even today. In this chapter we discuss the issues involved in creating an axiom system for Euclidean geometry from scratch.

Unit 11.1 Constructing Euclidean Geometry

In Section 7.1.1, we described **Euclidean geometry** as a mathematical system in which the statements that are assumed or can be deduced include all the axioms and propositions that are in Euclid's *Elements*. In trying to construct a rigorous development of Euclidean geometry, the first problem that faced mathematicians working in the era 1882–1910 was to choose undefined terms. Recall from Section 7.1.1 that Euclid had offered definitions for *all* important terms, so his work could not be used as historical precedent to help one decide which terms should be chosen. The only guide was that if a term T could be defined in terms of other terms $P_1, P_2, \ldots, P_n$ that were undefined or previously defined, then T did not need to be taken as undefined.

A quandary faced these mathematicians. How many terms needed to be undefined? Should geometry be separated from the rest of mathematics in the sense that no terms from outside the system (e.g., "number" or "function") would be assumed? Should undefined terms include logical terms such as "and" or "implies"? At this point in the history of mathematics, neither number nor logic had been placed on generally agreed-upon logical footings, so these mathematicians had major decisions to make.

Pasch chose *point, line segment, plane segment*, and *congruence of finite sets*. Peano worked from *point* and *line segment*. Pieri selected *point* and *rigid motion*. Veblen utilized *point* and *order*. Hilbert chose *point, line, plane, between, congruence of segments*, and *congruence of angles* as undefined terms in order to make his system a little more intuitive and easier to use.

In this and the next two sections we create a global system for Euclidean geometry. Our system is most like Hilbert's. We first establish our undefined terms as those pertaining to objects and those concerning relationships we associate with those objects. We take *point* and *line* as the basic *undefined objects* from which other objects of the geometry (segment, triangle, circle, angle, etc.) will be defined. Then we specify *undefined relations*, which, like our undefined objects, are implicitly defined through the axioms and will be used to define other relations. We want to be able to talk about a point being "on" a line or a circle (i.e., to specify whether a line or a circle contains that point). We should be able to identify some sort of order of points on a line, recognizing which points are "between" other points. Finally, we need to be able to compare the objects of our geometry by identifying which are equal or "congruent" to one another. To this end, we specify that *on, between*, and *congruent* are the undefined relations of our geometry, and these, along with the undefined objects, will be used to create the axioms and definitions we will need to prove theorems.

11.1.1 Axioms for incidence

We can classify the axioms that define Euclidean geometry into five categories: incidence, betweenness, congruence, continuity, and the parallel postulate. As you read, take the time to examine these axioms carefully, and use your intuition about the Euclidean geometry you studied in high school to convince yourself that they are indeed statements you can accept without proof. It is very helpful to draw diagrams to be sure that you understand exactly what each axiom allows you to say about the objects of geometry. We generally omit diagrams in this section to show that the propositions follow logically from the system without regard to any picture that may be drawn. We are careful only to use our undefined terms and relations (and any definitions derived from them) to create the axioms of our system. However, we do assume the familiar relations of sets and logic, including set membership, equality, implication, and the (natural) number of elements of a finite set.

Incidence axioms

The first axioms are called *incidence axioms* because they describe incidence properties of points and lines (i.e., what precisely it means for a point to be "on" a line, or for a line to be "on" a point). The relation of incidence is symmetric: If a point is incident with a line, then the line is incident with the point. In this case, we may say that "ℓ is on P" or "ℓ contains P" and that "P is on ℓ" or "P belongs to ℓ". The following incidence axioms for the Euclidean plane implicitly define the undefined relation *on*.

Axioms **Incidence:**

I-1: There exist at least three distinct points.

I-2: For each two distinct points there exists a unique line on both of them.

I-3: For every line there exist at least two distinct points on it.

I-4: Not all points lie on the same line.

Notice the importance of Axiom I-1. Without it, we do not know if we have any points.

Question 1: Why do Axioms I-1 and I-2 guarantee the existence of (at least) one line?

Axiom I-2 can be reformulated with the familiar statement: Two points determine a line. If A and B are points on line ℓ, we may write $\overleftrightarrow{AB}$ for ℓ, since they determine ℓ.

Notice that Axiom I-3 is not a definition of a line, but simply a statement telling us one of its properties. From this axiom alone, we cannot assume, for example, that a line is infinite, or even that it contains three points. For that, we need another axiom, which we introduce in Section 11.1.2.

Axioms I-1 and I-2 together imply that there exists at least one line. However, it is possible that our geometry consists only of that line. For us to establish that there is more than one line, we need Axioms I-3 and I-4.

What kinds of theorems of Euclidean geometry can be proved using only these incidence axioms? One theorem is the following.

Theorem 11.1 Given a point, there exist at least two distinct lines on it.

Proof: Let P be a point (I-1). By I-4, there exists at least one line ℓ not on P. There are at least two distinct points on ℓ (I-3), call them A and B. So $\overleftrightarrow{AB}$ does not contain P. By I-2, A and P determine a line. $\overleftrightarrow{AP}$ is not $\overleftrightarrow{AB}$ because P is not on $\overleftrightarrow{AB}$. Similarly, B and P determine a line that is not $\overleftrightarrow{AP}$ or $\overleftrightarrow{AB}$. Consequently, there are at least two lines on P. ┘

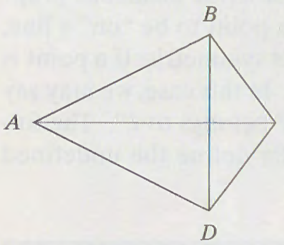
When two or more lines are on the same point, we say that these lines *intersect* in that point.

Geometries satisfying the incidence axioms that are not Euclidean

Axioms I-1 to I-4 are sufficient to prove there are at least three lines in the geometry (see Problem 3b). Also, these axioms appear intuitively true for Euclidean geometry. Can we now conclude that with *point, line,* and *on* as undefined terms, and with the incidence Axioms I-1 to I-4, we will be able to deduce all the theorems of Euclidean geometry? The answer is no. One way we can confirm this is to create a *model* (or interpretation) of these four axioms that is not Euclidean.

One such model has the four vertices of a tetrahedron $ABCD$ as its *points*, and the six edges of the tetrahedron $\{A, B\}$, $\{A, C\}$, $\{A, D\}$, $\{B, C\}$, $\{B, D\}$, and $\{C, D\}$ as its *lines* (Figure 1).

Figure 1



tetrahedron $ABCD$

(Convince yourself that Axioms I-1 to I-4 hold for this model.) Since for Euclidean geometry, we need infinitely many points, this model is not Euclidean.

Fano's geometry

Another model of Axioms I-1 to I-4 that is not Euclidean is called **Fano's plane geometry**, after the Italian mathematician Gino Fano (1871–1952).

Let **P** be the set of 7 points: $\{A, B, C, D, E, F, G\}$.

Let **L** be the set of 7 lines: $\{\{A, B, C\}, \{A, D, E\}, \{A, F, G\}, \{B, D, F\}, \{B, E, G\}, \{C, D, G\}, \{C, E, F\}\}$. We use set notation because no order to the points on a line is implied.

We now show that this *algebraic model* satisfies I-1 to I-4:

I-1: There exist at least three distinct points.
Since P contains 7 points, we can select any three to verify the axiom.

I-2: For each two distinct points there exists a unique line on both of them.
When we examine the lines we see that no two points belong to two different lines, and there is a line on every pair. To save space, we write ABC for the line $\{A, B, C\}$.

A, B ABC	B, C ABC	C, D CDG	D, E ADE	E, F CEF	F, G AFG
A, C ABC	B, D BDF	C, E CEF	D, F BDF	E, G BEG	
A, D ADE	B, E BEG	C, F CEF	D, G CDG		
A, E ADE	B, F BDF	C, G CDG			
A, F AFG	B, G BEG				
A, G AFG					

I-3: For every line there exist at least two distinct points on it.
This axiom is easily verified by observation. In fact, there are three distinct points on each line.

I-4: Not all points lie on the same line.
Since F does not belong to ABC, for example, the axiom is verified.

Question 2: Construct a geometric model of this system. Notice that there is nothing in the axioms that imply a line must be drawn straight! One geometric model of this system can be constructed using a "triangle" with an inscribed "circle".

These two models, the first consisting of 4 points and the second with 7 points, are examples of what are called *finite incidence geometries*. The study of finite incidence geometries was initiated by Fano in 1892, when he introduced a three-dimensional geometry with 15 points and 35 lines, in which each plane is on 7 points.

Question 3: What is the smallest finite geometry that can be created from the incidence axioms? Give an algebraic and geometric model.

All the models introduced in this section, as well as Euclidean geometry, are classified under the general title of **incidence geometries**, i.e., geometries consisting of points and lines with the relation *on*, and that satisfy Axioms I-1 to I-4.

These models show that Axioms I-1 to I-4 are not sufficient to prove that a line contains infinitely many points. We know this is a property of Euclidean lines. In the next section we introduce the required axioms to "fill in" the gaps between and beyond the points of the line given by the incidence axioms.

11.1.1 Problems

1. Create a model of a 3-point geometry that satisfies Axioms I-1 to I-4.

2. a. Create a model of a finite geometry that satisfies only two of Axioms I-1 to I-4.

b. Create a model of a finite geometry that satisfies only three of Axioms I-1 to I-4.

3. Prove the following theorems of Euclidean geometry using the incidence axioms alone.

a. The intersection of two distinct lines is exactly one point.

b. There exist three nonconcurrent lines (i.e., lines that are not on one common point).

c. For every line there is at least one point not on it.

4. Prove the following theorems of Euclidean geometry using the incidence axioms.

a. If a point A is not on a line ℓ containing distinct points B and C, then there is no line containing all three of them.

b. If there is no line containing all three points A, B, and C, then A, B, and C are distinct points.

5. Consider the following axioms, based on the undefined terms *point, line*, and *on*:

F-1: There exists at least one line.

F-2: There are exactly three points on every line.

F-3: Not all points are on the same line.

F-4: There exists exactly one line on any two distinct points.

F-5: There exists at least one point on any two distinct lines.

Show that Fano's 7-point geometry satisfies these axioms.

6. Rewrite the axioms for Fano's plane geometry (Problem 5) interchanging the words "point" and "line". Each of these new axioms is called the **dual** of its counterpart. Create a model for this new axiom set and determine if it also satisfies F-1 through F-5.

7. See Problem 5.

a. Construct a model for a geometry that satisfies F-1 and F-2, but not F-3.

b. Construct another model for a geometry that satisfies F-1 and F-3, but not F-2.

c. Determine whether each of your models is a model of incidence geometry. Explain why or why not.

8. a. Using the axioms for Fano's plane geometry (Problem 5), prove that each point is on exactly three lines.

b. Use Fano's model to answer the following question: A Rotary club with seven new members wants to plan meetings for the new members so they can get to know one another. The club wants each new member to meet with every other new member exactly once, and to have exactly three new members present at each meeting. How can such meetings be arranged, and how many of these meetings will each new member attend?

9. Let **P** be the set of 9 points $\{A, B, C, D, E, F, G, H, I\}$. Let **L** be the set of 12 lines $\{ABC, DEF, GHI, ADG, BEH, CFI, AEI, BFG, CDH, AFH, BDI, CEG\}$. Here, as in this section, we write ABC for $\{A, B, C\}$, etc. Show that this system, with the relation "on", satisfies axioms I-1 to I-4. This 9-point geometry is called **Young's geometry**. It is named after the famous mathematician John Wesley Young (1879–1932), who discovered it.

10. Consider the following model. The set of points is the interior of a given Euclidean circle. The set of lines consists of the open chords of the circle (an open chord is a chord minus its endpoints). Explain why this model satisfies or does not satisfy Axioms I-1 to I-4.

11. Consider the model in which the set of points are those inside a given Euclidean triangle, and the set of lines consists of the open segments joining two points lying on different sides of the triangle (an *open segment* is a segment without its endpoints). Explain why this model satisfies or does not satisfy Axioms I-1 to I-4.

12. Consider the follow axiom set, based on the undefined terms *point, line*, and *on*.

Axiom 1: There are exactly four points.

Axiom 2: For each two distinct points there is a unique line on them.

Axiom 3: There are exactly two points on every line.

Construct a model for this geometry and create and prove one theorem for it.

13. **Desargues's Theorem**, named after the French mathematician Girard Desargues (1596–1660), states that if ΔABC and ΔDEF are so situated that the lines $\overleftrightarrow{AD}$, $\overleftrightarrow{BE}$, $\overleftrightarrow{CF}$ all meet in a point, then the intersections of sides $\overleftrightarrow{AB}$ and $\overleftrightarrow{DE}$, $\overleftrightarrow{BC}$ and $\overleftrightarrow{EF}$, $\overleftrightarrow{CA}$ and $\overleftrightarrow{FD}$ are on the same line.

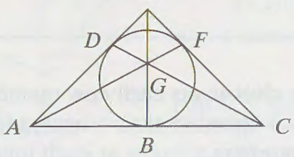
a. Draw a picture of this theorem in the Euclidean plane, choosing the vertices of ΔABC and ΔDEF so that no pair of corresponding sides is parallel.

b. This theorem is true in Fano's plane geometry of 7 points and 7 lines. Create two triangles in Fano's geometry (assuming a triangle is simply a set of three points not on the same line) that satisfy the hypothesis of Desargues's Theorem, and show that the conclusion holds.

ANSWERS TO QUESTIONS

1. There exist three distinct points by Axiom I-1. Call them A, B, C. Then, by Axiom I-2, there is at least one line $\overleftrightarrow{AB}$.
2. See Figure 2.

Figure 2



3. I-3 tells us that there exist at least three distinct points. Call them A, B, and C. By I-1 there exists a unique line on each pair of them. So there are at least three lines: $\overleftrightarrow{AB}$, $\overleftrightarrow{AC}$, and $\overleftrightarrow{BC}$. These lines satisfy the conditions of I-2: For every line there exist at least two distinct points on it. And I-4 is satisfied because B is not on $\overleftrightarrow{AC}$. So the smallest plane incidence geometry consists of the set of points $P = \{A, B, C\}$ and the set of lines $L = \{\overleftrightarrow{AB}, \overleftrightarrow{AC}, \overleftrightarrow{BC}\}$. A geometric model of this geometry consists of the vertices of triangle ABC.

11.1.2 Axioms for betweenness

We have seen that the incidence axioms alone do not imply that there are infinitely many points on a line. To achieve this, as well as to enable us to "order" points on a line, we introduce *betweenness* of points, and we specify the set of axioms that reveal the properties of this concept.

The betweenness axioms

Suppose a line contains four points A, B, C, and D. Assume A is between B and C, and C is between A and D. What can be concluded about A, B, and D? Suppose these four points were on a circle. Would the conclusion be the same? Euclid's axioms did not provide the means to determine the answers to these questions. The axioms that follow do. They also ensure that a geometry that satisfies them and the incidence axioms will contain infinitely many points. Notice that, just as with the incidence axioms, the only objects and relations we use are those we have accepted as undefined terms. You should draw diagrams to be sure you understand the properties of points that these axioms reveal.

Axiom B-1 is like Euclid's second postulate, which guarantees that a line can be extended "continuously". Axioms B-2 through B-4 ensure that there is an order to points on a line. For this reason some people call these "axioms of order". These axioms also enable us to distinguish lines from closed curves.

Axioms | **Betweenness:**

B-1: Let A and B be two distinct points. There exist points C, D, and E on line $\overleftrightarrow{AB}$ such that C is between A and B, B is between A and D, and A is between E and B.

B-2: If A, B, and C are points such that B is between A and C, then A, B, and C are distinct and on the same line.

B-3: If A, B, and C are points such that B is between A and C, then B is between C and A.

B-4: If A, B, and C are three distinct points on the same line, then exactly one of the following statements is true: B is between A and C; C is between A and B; or A is between C and B.

Question 1: What do Axiom B-1 and Axiom I-3 have in common?

Axiom B-2 makes the necessary connection between the incidence axioms and the betweenness axioms, relating the concept of *betweenness* to the *incidence* properties of point and line. Without it, we would have two separate theories—one for incidence, and one for betweenness.

Question 2: Axiom I-1 guarantees that there exist at least three distinct points. How many points do we obtain by Axioms B-1 and B-2?

From now on, because of Axioms B-1 and B-2, we do not need to be concerned that there are an insufficient number of points necessary for our theorems. Axiom B-3 says we can symmetrically permute A and C in the relation "B is between A and C", without destroying its validity, but Axiom B-4 says that if we apply certain other cyclic permutations, the validity of the relation is destroyed. Axiom B-4 indicates that all points on a line can be lined up so that, given any three points, one is between the other two.

At this juncture we are able to make certain definitions that would not have been possible before. It is important to realize that definitions are meaningful in a mathematical theory only if the axioms imply that the object being defined actually exists. With B-1 to B-4, for example, we are able to define *segment* because these betweenness axioms indicate that there are points between two given points, and others that are not. We need the concept of segment before we introduce our next betweenness axiom.

Definitions

Let A and B be two distinct points. **Segment AB**, written $\overline{\mathbf{AB}}$, is the set consisting of A and B and all points on the line $\overleftrightarrow{AB}$ that are between A and B. A and B are called the **endpoints** of $\overline{AB}$.

Notice we have defined segment using only the undefined terms of our system: "point", "line", "on", and "between".

Now we are free to incorporate the idea of segment into the definitions and axioms that follow. One idea that you have seen on occasion and that is defined in terms of segments is *convexity*. For instance, we spoke of convex and nonconvex kytes in Section 7.3.1 (Problem 9). Recall that a set S of points is **convex** if and only if whenever A and B are in S, so are all points between A and B. Since many elementary geometry books speak little if at all about convexity, many people think convexity is a frill or nonessential part of the study of geometry. Quite the contrary. A last betweenness assumption, Axiom B-5, involves convexity.

We can define the **plane** to be the set of all points and lines in this geometry. Our last betweenness axiom allows a plane to be partitioned into a line and two other parts, and is therefore described as a "plane separation" axiom.

Axiom

Plane Separation:

B-5: Every line ℓ partitions the plane into ℓ itself and two convex sets S_1 and S_2 such that if a point P is in S_1 and a point Q is in S_2, then $\overline{PQ}$ and ℓ have a point in common.

We say that S_1 and S_2 are the **sides** or **half-planes** of the line ℓ, and that they are **bounded** by ℓ. It is also common to say that S_1 and S_2 are **opposite sides of** ℓ. Recall the invalid "proof" in Section 7.1.2 that every triangle is isosceles. Our explanation of the fallacy relied on a careful definition of sides of a line that the plane separation axiom and the above definitions provide.

The **triangle with vertices A, B, C,** denoted **ΔABC**, is defined to be the union of the segments $\overline{AB}$, $\overline{BC}$, and $\overline{AC}$, its **sides**. With Axiom B-5, we are able to prove a statement based on an axiom introduced by Pasch in 1882. This statement essentially states that a line that goes into a triangle must come out.

Theorem 11.2 **(Postulate of Pasch):** If line ℓ different from $\overleftrightarrow{AB}$ intersects side $\overline{AB}$ of ΔABC in a point between A and B, then exactly one of the following holds: (1) ℓ contains C, (2) ℓ contains a point on $\overline{AC}$ between A and C, or (3) ℓ contains a point on $\overline{BC}$ between B and C.

Proof: Let line ℓ, different from $\overleftrightarrow{AB}$, intersect side $\overline{AB}$ in a point X between A and B. Then ℓ cannot also contain A or B because it would then equal $\overleftrightarrow{AB}$. Since ℓ intersects $\overleftrightarrow{AB}$ between A and B, A and B are on opposite sides of ℓ. Now, by the Plane Separation Axiom, either (1) C is on ℓ; (2) C is on the same side of ℓ as A, in which case C is on the opposite side of ℓ as B and so ℓ intersects $\overline{BC}$; or (3) C is on the same side of ℓ as B, in which case C is on the opposite side of ℓ as A, and so ℓ intersects $\overline{AC}$. ∎

Theorem 11.2 fills a significant gap in the deductive reasoning of the *Elements*. Euclid relied in many theorems on visual clues from diagrams, instead of using Pasch's axiom (or one equivalent to it) to validate his proofs.

The Plane Separation Axiom also allows us to order more than three points on a line. Theorem 11.3 and its corollary allow us to establish order for four points on a line. To help understand this theorem and its corollary, we introduce some notation. We write **A-B-C** if and only if B is between A and C.

Theorem 11.3 Let A, B, and C be points on a line with A-B-C. Suppose D is a fourth point on the line such that A-C-D. Then A-B-D and B-C-D.

Proof: Let A, B, C, and D be four distinct points on line ℓ such that A-B-C and A-C-D. There exists a point E not on ℓ. (Why?) Let $m = \overleftrightarrow{EC}$. Then $m \neq \ell$ (because E is on m and not on ℓ) and A, B, and D are not on m (because otherwise $m = \ell$). So $AD \neq m$. (Why?) Since A-C-D, $\overline{AD}$ intersects m at C. Consequently, A and D are on opposite sides of m. Now since A-B-C, $\overline{AB}$ does not intersect m (otherwise C is between A and B). Thus, by Plane Separation, A and B are on the same side of m. Consequently, again by Plane Separation, B and D are on opposite sides of m, and since $\overline{BD}$ intersects m at C, B-C-D.

To show that A-B-D, let $n = \overleftrightarrow{EB}$ and proceed in the same way as with m. We leave the details to you. ∎

A similar proof leads to this corollary. See Problem 4.

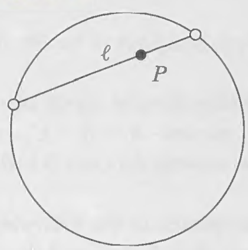
Corollary: (1): If A-B-C and B-C-D, then A-B-D and A-C-D.
(2): If A-B-D and B-C-D, then A-B-C and A-C-D.

When points A, B, C, and D satisfy a given condition of the corollary, we can write **A-B-C-D** because we can delete any one of these 4 points and the rest are in the same order.

The betweenness axioms provide us with the infinite lines and order of points on lines we need for Euclidean geometry. Does that mean that these axioms, in conjunction with the incidence axioms, are sufficient to define Euclidean geometry? Again, we can demonstrate they are not by constructing a model of these axioms that is not Euclidean.

Consider the *Beltrami-Cayley-Klein plane* model, which is named for Eugenio Beltrami (1835–1900), Arthur Cayley (1821–1895), and Felix Klein (1849–1929). Beltrami and Klein independently discovered it in 1871. However, it also appeared in an 1859 paper by Cayley.

Figure 3



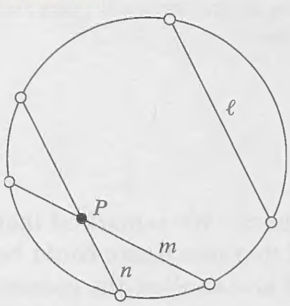
In the **Beltrami-Cayley-Klein plane model**, a point is interpreted as a point interior to a Euclidean circle **C**. A line is any open chord of **C**, i.e., a chord of the circle without its endpoints. Figure 3 shows a point P on a line ℓ in this model. Since the environment of this geometry is the interior of a Euclidean circle, it would appear that this could be a model of Euclidean geometry. We know that the incidence and betweenness axioms hold for Euclidean points and lines. Since chords can be viewed as subsets of lines, it makes intuitive sense that these axioms also hold in this model (see Problem 10 of Section 11.1.1).

Consider, for example, Axiom I-2. To show it holds in the Beltrami-Cayley-Klein plane, we need to show that for any two distinct points A and B in the interior of **C** there is a unique open chord containing them.

Proof of Axiom I-2 for the Beltrami-Cayley-Klein plane: Let A and B be interior to **C**. Let $\overleftrightarrow{AB}$ be the unique Euclidean line on them by Axiom I-2. This line intersects the boundary of **C** in two points. (This is a theorem of Euclidean geometry that ensures a line passing through the interior of a circle will intersect the circle in two distinct points. We discuss it in Section 11.1.3.) Call these points C and D. Then A and B lie on the open chord CD, which by Axiom I-1 for Euclidean geometry, is the only open chord on which they both lie. ∎

The other incidence axioms and betweenness axioms can be likewise verified for this model (see Problems 5 and 6).

Figure 4



$m \parallel \ell, n \parallel \ell$

Once the incidence and betweenness axioms are verified, does this mean that the Beltrami-Cayley-Klein plane is a model of Euclidean geometry? No. It turns out that the geometry this model describes is not Euclidean! Here is one reason why (there are many others!): In Euclidean geometry there is only one line parallel to a given line through a point outside of that line. (By "parallel" we mean "nonintersecting".) We can see from Figure 4 that it is possible to have at least two such lines in the Beltrami-Cayley-Klein plane. In fact, there are infinitely many lines parallel to ℓ through P.

The Beltrami-Cayley-Klein plane is a model of a *non-Euclidean geometry*. Such geometries are infinite incidence geometries that share certain Euclidean axioms but that replace the Euclidean parallel postulate, which we discuss in Section 11.1.4, with an axiom of multiple parallels, or with an axiom that denies the existence of parallels. The Beltrami-Cayley-Klein plane is a model of the non-Euclidean geometry that is called *hyperbolic* or *Bolyai-Lobachevskian*, named for the Russian mathematician Nikolai Ivanovich Lobachevsky (1792–1856) and the Hungarian mathematician Jànos Bolyai (1802–1860), who first discovered it.

In hyperbolic geometry, the incidence and betweenness axioms I-1 to I-4 and B-1 to B-5 hold. For this reason, hyperbolic geometry is considered, as is Euclidean geometry, an *ordered incidence geometry*. But there must be certain axioms that distinguish hyperbolic from Euclidean geometry since it is non-Euclidean. From the discussion above, we can see that the Euclidean parallel postulate is false in hyperbolic geometry. Are there other Euclidean propositions that do not hold in hyperbolic geometry? Yes, in hyperbolic geometry, if two triangles are similar, then they are congruent. We next explore the congruence axioms of Euclidean geometry to determine if these are sufficient to categorize it, and perhaps further distance it (or not!) from hyperbolic geometry.

11.1.2 Problems

1. a. Which betweenness axiom guarantees that $\overline{AB}$ contains at least three points?

b. Explain why $\overline{AB}$ represents the same set of points as $\overline{BA}$.

2. Rewrite Axioms B1 to B4 using the *A-B-C* notation for betweenness of points.

3. Complete the proof of Theorem 11.3.

4. Prove the Corollary to Theorem 11.3.

5. Verify incidence Axioms I-1, I-3, and I-4 in the Beltrami-Cayley-Klein model of hyperbolic geometry.

6. Consider the following analytic description of the Beltrami-Cayley-Klein plane: A point is an ordered pair (x, y) such that x and y are real numbers and $x^2 + y^2 < 1$. A line is any nonempty set of points (x, y) that satisfies some equation $ax + by + c = 0$, such that a and b are real numbers not both zero. Verify betweenness Axioms B-2 to B-5 for this model.

7. Let ℓ be an arbitrary line in a given plane. Prove that ℓ determines exactly two distinct half-planes.

8. Look up Euclid's proof of his Proposition 21 in Book I. Determine where in the proof there is a statement that requires Pasch's axiom.

9. Interpret points to be real numbers, and lines to be sets of real numbers.

a. If the betweenness relation is interpreted in terms of $<$ (i.e., B is between A and C means $A < B < C$, or $C < B < A$), show that the betweenness Axioms B1–B4 are verified.

b. If the betweenness relation is interpreted in the following way: B is between A and C means that $A \neq C$ and there exist positive real numbers x and y such that $B = Ax + Cy$, where $x + y = 1$, show the betweenness Axioms B1–B4 are verified.

10. Can the betweenness axioms hold in a finite incidence geometry? Support your answer by either showing they are verified in one of the finite incidence geometries discussed in Section 11.1.1, or explaining why they do not.

ANSWERS TO QUESTIONS

1. They are both existence axioms.

2. These axioms provide us with the means to generate an infinity of points on a line. We know by Axiom I-1 there exist three distinct points. Call them A, B, and C. By Axiom B-1, for distinct points A and B, we generate additional points D between A and B, E such that B is between A and E, and F, such that A is between F and B. These points are distinct from each other and from A and B by Axiom B-2. Now we can continue this process choosing distinct points in pairs and generating other points on the line.

11.1.3 Congruence and the basic figures

In Chapter 7, we detailed Euclid's treatment of congruence. We remarked that Euclid's approach to congruence assumed without proof that one figure could be superimposed onto another. To correct that weakness and also to allow the geometry of congruence to apply to all figures, we showed how congruence can be developed through transformations. We used synthetic, coordinate, and complex number descriptions of these transformations to deduce properties of congruence. Clearly, along the way we were utilizing a great deal of knowledge about Euclidean geometry, about functions, and about number.

A pure synthetic approach

We called one of our earlier approaches *synthetic* because it did not use coordinates or complex numbers. However, we did utilize numbers for distance and angle measure, and we defined congruence in terms of distance. So it was not a *pure* synthetic approach to congruence. The use of numbers in an approach which is otherwise synthetic is the most common current practice, and it is quite efficient and effective because it allows everything we know about numbers to be used. However, it leads many students to believe that geometry cannot be divorced from numbers.

A pure synthetic approach does not allow numbers at all. In this section we exhibit such an approach. We start from scratch about congruence, assuming only what we have explicitly mentioned in Sections 11.1.1 and 11.1.2. These axioms, together with the congruence axioms we introduce in this section, correct the defects in Euclid's approach to congruence and enable all his theorems about congruence to be deduced. Our approach here also shows how some terms previously defined in terms of number (e.g., supplementary angles and midpoints of segments) can be defined without number. We show how to examine two angles or two segments and determine that one is bigger than another without using numbers. We also treat triangle congruence without using numbers. Along the way, you will see many definitions of terms that you have been using in earlier chapters. Some are the same as in previous chapters and are repeated here so that this unit can be self-contained. Others are different due to the approach we are taking but are still not contradictory to definitions in earlier chapters.

Congruence axioms

Remember, in addition to *on* and *between*, we take the relation *is congruent to* (denoted "$\cong$") as an undefined term. In the following definitions and axioms, we develop the properties of congruent segments and angles. We first need to introduce the concept of *ray*. These definitions are the familiar ones.

Definitions

Given distinct points A and B, **ray AB**, written $\overrightarrow{\mathbf{AB}}$, is the set of points of the segment $\overline{AB}$ together with all points C on line $\overleftrightarrow{AB}$ such that B is between A and C. A is called the **vertex** (or **endpoint**) of the ray. $\overrightarrow{\mathbf{AB}}$ and $\overrightarrow{\mathbf{AC}}$ are **opposite rays** if A is between B and C.

Like a segment, a ray is convex. And, from their definitions, both segments and rays are subsets of the lines containing them. Rays can therefore be related to half-planes, or sides of a given line in the following way: If A is a point on a line ℓ and B is a point not on ℓ, then ray $\overrightarrow{AB}$ includes point A and exactly those points on line $\overleftrightarrow{AB}$ on the same side (half-plane) of ℓ as B. The only other points on $\overleftrightarrow{AB}$ are the points other than A on the ray $\overrightarrow{AC}$ opposite ray $\overrightarrow{AB}$. These points are on the opposite side of ℓ as B. We say that line ℓ **separates** line $\overleftrightarrow{AB}$ at A.

Segment congruence

Three axioms treat congruence of segments. Again you should draw diagrams to convince yourself that these axioms make sense to you.

Axioms

Segment Congruence:

C-1: Let A and B be distinct points. If C is any point, then for each ray r with vertex C, there exists a unique point D on r such that D is distinct from C and $\overline{AB}$ is congruent to $\overline{CD}$, written $\overline{AB} \cong \overline{CD}$.

C-2: Segment congruence is reflexive, symmetric, and transitive. That is, segment congruence is an equivalence relation.

C-3: Let B be between A and C, and E be between D and F. If $\overline{AB} \cong \overline{DE}$ and $\overline{BC} \cong \overline{EF}$, then $\overline{AC} \cong \overline{DF}$.

Question 1: Which of Euclid's common notions, if any, do each of the congruence axioms C-1, C-2, and C-3 replace?

Axiom C-1 gives justification for the familiar operation of "laying off" a segment on a ray. Intuitively, it tells us we can "move" the segment along the ray. Axiom C-3, in effect, tells us that if we "add" congruent segments, the sums are congruent. Notice that we are able to specify this axiom without introducing the concept of distance or length. From it we are able to prove the corresponding "segment subtraction" property.

Theorem 11.4 Let B be between A and C, and E be between D and F. If $\overline{AC} \cong \overline{DF}$ and $\overline{AB} \cong \overline{DE}$, then $\overline{BC} \cong \overline{EF}$.

Proof: Using Axiom C-1, let C' be the point on ray $\overrightarrow{BC}$ such that $\overline{BC'} \cong \overline{EF}$. Using this and the given $\overline{AB} \cong \overline{DE}$, $\overline{AC'} \cong \overline{DF}$ by Axiom C-3. But $\overline{AC} \cong \overline{DF}$ is given. So by Axiom C-2, $\overline{AC} \cong \overline{AC'}$. Since A-B-C, C is on $\overrightarrow{AB}$. We would like to show that $C = C'$. For this, we need to show that C' is on $\overrightarrow{AB}$. Since C is on $\overrightarrow{BC}$ either (1) B-C-C', (2) B-C'-C, or (3) $C = C'$. (1) If B-C-C', then by Theorem 11.3, A-B-C', so C' is on $\overrightarrow{AB}$; (2) if B-C'-C, then by the Corollary to Theorem 11.3, A-B-C', so C' is on $\overrightarrow{AB}$. So in all cases C' is on $\overrightarrow{AB}$. Because C and C' are on the ray $\overrightarrow{AB}$, by Axiom C-1, $C = C'$. Consequently, $\overline{BC} \cong \overline{BC'}$, and so using Axiom C-2 once again, $\overline{BC} \cong \overline{EF}$. ∎

With a suitable definition of "greater than", we can order segments by their size even without numbers!

Definition **Segment $\overline{CD}$ is greater than segment $\overline{AB}$**, written $\overline{CD} > \overline{AB}$, if and only if there exists a point E between C and D such that $\overline{CE} \cong \overline{AB}$.

We can now prove an important order relation for segments.

Theorem 11.5 If $\overline{CD} > \overline{AB}$ and $\overline{AB} \cong \overline{EF}$, then $\overline{CD} > \overline{EF}$.

Proof: If $\overline{CD} > \overline{AB}$, then by definition, there exists a point X between C and D such that $\overline{CX} \cong \overline{AB}$. If $\overline{AB} \cong \overline{EF}$ and $\overline{CX} \cong \overline{AB}$, then by Axiom C-2, $\overline{CX} \cong \overline{EF}$. Thus X is a point between C and D such that $\overline{CX} \cong \overline{EF}$. Therefore, by definition, $\overline{CD} > \overline{EF}$. ∎

We ask you to prove other order relations for segments in Problem 11.

With Axioms C-1 to C-3, we can now define some other important objects of Euclidean geometry in terms of congruence, and distance and length are not involved.

Definitions Let O and A be distinct points. The set of all points B such that $\overline{OB} \cong \overline{OA}$ is the **circle** with center O. Each of the segments $\overline{OB}$ is a **radius** of the circle.

Recall that Euclid explicitly postulated the existence of a circle (see Section 7.1.1), instead of making it the subject of a definition as we have done.

Question 2: How do we know the points B of the definition exist? How many of these points are necessary to determine a circle completely?

It turns out that although such points B exist, there are not yet sufficient axioms to actually prove that given three noncollinear points, a circle exists that contains them. You may be surprised to learn that the Euclidean parallel postulate is needed to prove this.

Angle congruence

From segments we proceed to angles. The concept of betweenness of points leads in a natural way to the concept of betweenness of rays, and that to the concept of angle. We can make the analogy that betweenness of rays is related to angles as betweenness of points is related to segments.

Definition Let $\overrightarrow{AB}$, $\overrightarrow{AC}$, $\overrightarrow{AD}$ be distinct rays such that $\overrightarrow{AB}$ and $\overrightarrow{AD}$ are not opposite. Ray $\overrightarrow{AC}$ **is between rays** $\overrightarrow{AB}$ **and** $\overrightarrow{AD}$ if and only if there exist points X, Y, and Z, such that X belongs to $\overrightarrow{AB}$, Y belongs to $\overrightarrow{AC}$, Z belongs to $\overrightarrow{AD}$, and Y is between X and Z.

Because betweenness of rays is defined in terms of betweenness of points, each betweenness axiom for points B1–B4 has a counterpart in the betweenness of rays.

Question 3: What betweenness axiom allows us to conclude that betweenness of rays is symmetric?

It is also possible to prove that if $\overrightarrow{AC}$ is between $\overrightarrow{AB}$ and $\overrightarrow{AD}$, then, other than A,

i. each point of $\overrightarrow{AB}$ and each point of $\overrightarrow{AC}$ are on the same side of $\overleftrightarrow{AD}$, and likewise,
ii. each point of $\overrightarrow{AC}$ and each point of $\overrightarrow{AD}$ are on the same side of $\overleftrightarrow{AB}$,
iii. each point of $\overrightarrow{AB}$ and each point of $\overrightarrow{AD}$ are on opposite sides of $\overleftrightarrow{AC}$.

With the correspondence between betweenness of points and betweenness of rays in mind, we can use many of the results we achieved for betweenness of points to prove theorems about betweenness of rays.

Question 4: Can the following property of betweenness of points be extended to rays? If B is between A and C and C is between B and D, then B is between A and D.

Angles

We now turn to angles. The following definition should be familiar.

Definitions An **angle** (written $\angle$) is the union of two distinct and nonopposite rays $\overrightarrow{AB}$ and $\overrightarrow{AC}$, called its **sides**. A is the **vertex** of the angle. The angle is denoted as $\angle BAC$ or $\angle CAB$. A point D is in the **interior of $\angle$BAC** if D is on the same side of $\overleftrightarrow{AC}$ as B and if D is also on the same side of $\overleftrightarrow{AB}$ as C.

Question 5: Describe the interior of an angle in terms of betweenness of rays and in terms of half planes.

It turns out that just as we can separate a plane by a line, we can separate a plane by an angle, into those points that are in its interior and those that are not.

Some developments of geometry allow *straight angles*, defined as angles that are the union of opposite rays, angles that in an analytic approach have measure 180°. Such angles, when they are allowed, can cause difficulties because they are identical to lines, any point on them can be a vertex, and either side of the line can be the interior of the angle. When Euclid wished to speak of the equivalent of 180°, he used the phrase "two right angles", as you can see from Proposition 32 in the list in Section 7.1.1. Some developments also allow *zero angles*, defined as angles that are the union of identical rays, angles that in an analytic approach have measure 0°. Such an angle would have no points in its interior and would be identical to a single ray. And we have pointed out in Section 7.2.2 that some books speak of *reflex angles*, angles with measures greater than 180°. Allowing reflex angles requires that the union of two rays determines two angles. All of these ideas are possible in a rigorous treatment of geometry, but because they tend to complicate the development, we do not employ them here.

The following axioms are the assumed properties of congruent angles. Notice that they do for angles what Axioms C-1, C-2, and C-3 do for segments. Axioms C-1 and C-4 convey the analytic idea that segments and angles have exactly one measure. In an analytic approach, we could replace Axioms C-2 and C-5 by assuming that congruent segments have the same measure and congruent angles have the same measure. Axioms C-3 and C-6 tell us that if we put together adjacent congruent segments and angles, the results are congruent.

Axioms

Angle Congruence:

C-4: Given $\angle BAC$ and ray $\overrightarrow{DE}$, there exist unique rays $\overrightarrow{DF}$ and $\overrightarrow{DG}$ on different sides of $\overleftrightarrow{DE}$ such that $\angle BAC \cong \angle EDF \cong \angle EDG$.

C-5: Angle congruence is reflexive, symmetric, and transitive. That is, angle congruence is an equivalence relation.

C-6: Let D be in the interior of $\angle ABC$ and E be in the interior of $\angle GHI$. If $\angle ABD \cong \angle GHE$ and $\angle DBC \cong \angle EHI$, then $\angle ABC \cong \angle GHI$.

From Axioms C-4 to C-6, we can deduce "angle subtraction" in a manner similar to the proof of Theorem 11.4.

Theorem 11.6 Let D be in the interior of $\angle ABC$ and E be in the interior of $\angle GHI$. If $\angle ABC \cong \angle GHI$ and $\angle ABD \cong \angle GHE$, then $\angle DBC \cong \angle EHI$.

Proof: The proof is left to you as Problem 12. ∎

The next definition enables us to be able to compare angles that are *not* congruent. It is analogous to the earlier definition in this section comparing segments.

Definition $\angle DEF$ **is greater than** $\angle ABC$, written $\angle DEF > \angle ABC$, if and only if there exists a ray $\overrightarrow{EG}$ such that $\angle ABC \cong \angle DEG$ and G is in the interior of $\angle DEF$.

Terms defined in earlier chapters without reference to number can be used here. Two angles with a common vertex form a pair of **vertical angles** if the sides of one are opposite to the sides of the other. Two angles form a **linear pair** if they have a common side and their noncommon sides are opposite rays. Two angles are **adjacent angles** if they are a linear pair, or if they have a common vertex and a common side that is in the interior of the angle formed by their noncommon sides.

These definitions enable us to define supplementary angles and right angles using congruence and with no reference to angle measure or other numbers. So these definitions are likely to be different from those you have seen before.

Definitions Two angles are **supplementary** and each is a **supplement** of the other if they are respectively congruent to the angles of a linear pair.

A **right angle** is an angle that is congruent to a supplement of itself.

Triangle congruence

Thus far, all of the congruence axioms have dealt with relationships along a line or around a point. We need an axiom which guarantees that congruence in one part of the plane is like congruence in every other part. An axiom that fills this void is *Side-Angle-Side (SAS) congruence for triangles.*

Axiom **Triangle Congruence Axiom (SAS Congruence):**

C-7: Given two triangles, if two sides of one are congruent to two sides of the other, and the angles included by the congruent sides are congruent, then their third sides are congruent and their remaining corresponding angles are congruent.

When all six parts (sides and angles) of one triangle are congruent to the corresponding parts of another triangle, we call them **congruent triangles**. Thus, triangle congruence is defined in terms of the undefined terms *segment congruence* and *angle congruence*, and the properties of triangle congruence are known only by what can be deduced from postulates C-1 to C-7.

It follows from the definition of triangle congruence and Axioms C-2 and C-5 that triangle congruence is reflexive, symmetric, and transitive.

Notice that the statement of the SAS Congruence Axiom C-7 is very much like Euclid's statement of the SAS proposition. It does not mention congruent triangles! Rather, it goes directly to the usual reason for proving triangles congruent, namely, to obtain the congruence of the other sides and the other two pairs of corresponding angles.

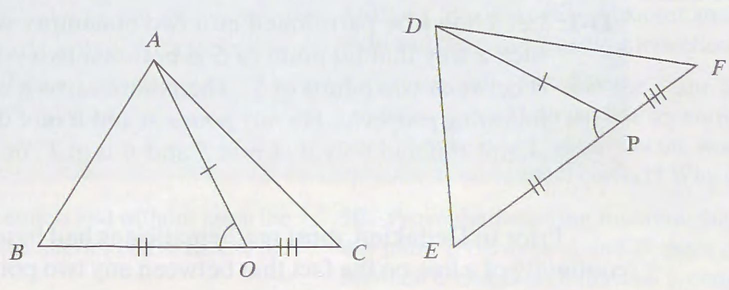
The SAS Congruence Axiom is very powerful—it connects congruence of angle and congruence of segments, which until this point have, in a sense, been two separate theories. It also enables us to deduce all the other triangle congruence propositions.

You can see how the connection is made between congruent segments and congruent angles in the following theorem. Although its proof seems long, it is quite straightforward, simply involving repeated use of the SAS Congruence Axiom to achieve the result.

Theorem 11.7 Supplements of congruent angles are congruent.

Proof: It is sufficient to consider the case where the pairs of supplementary angles are linear pairs, since (by Axiom C-5) congruence of angles is an equivalence relation. Suppose then that $\angle AOB$ and $\angle AOC$, and $\angle DPE$ and $\angle DPF$ are two linear pairs, and that $\angle AOB \cong \angle DPE$. We show that $\angle AOC$ and $\angle DPF$ are congruent. We can assume that segments $\overline{OA}$, $\overline{OB}$, and $\overline{OC}$ are congruent respectively to $\overline{PD}$, $\overline{PE}$, and $\overline{PF}$ (Axiom C-1). (See Figure 5.) Then $\overline{AB} \cong \overline{DE}$ and $\angle OBA \cong \angle PED$ by SAS Congruence (Axiom C-7). Furthermore, we know that O is between B and C and that P is between E and F (definition of opposite rays). So $\overline{BC} \cong \overline{EF}$ by Axioms C-2 and C-3. Since $\angle ABC \cong \angle DEF$, then again by SAS Congruence, $\overline{AC} \cong \overline{DF}$ and $\angle ACB \cong \angle DFE$. Thus $\angle AOC \cong \angle DPF$ by SAS Congruence. ∎

Figure 5



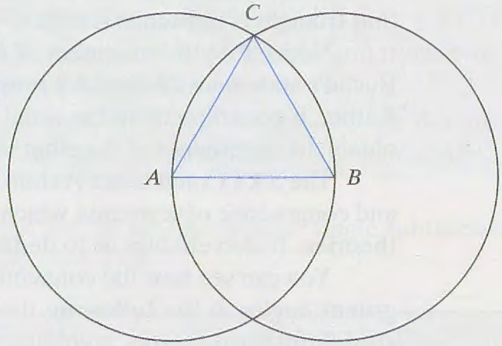
The need for an additional axiom

The incidence, betweenness, and congruence axioms allow the proving of many theorems about circles and polygons. But they do not enable rigorous proofs of some of the basic theorems that are in Euclid's geometry. As an example, consider Euclid's very first theorem and the proof he gave.

Euclid's Theorem 1 An equilateral triangle can be constructed on any segment $\overline{AB}$.

Proof: Draw a circle with radius $\overline{AB}$ and center A, and draw the circle with radius $\overline{AB}$ and center B (see Figure 6). (These can be drawn because of Euclid's Postulate 3.) Let C be a point of intersection of the circles. Now $\overline{AC} \cong \overline{AB}$ (Euclid's definition of "circle" applied to the circle with center A) and $\overline{BC} \cong \overline{AB}$ (Euclid's definition of "circle" applied to the circle with center B). By Common Notion 1, $\overline{AC} \cong \overline{BC}$. And so all sides of triangle ABC are congruent to each other, and the triangle is therefore equilateral. ∎

Figure 6



(We note that all of Euclid's proofs are written in paragraph form. The two-column form for writing proofs is a relatively recent invention, traceable back only to the 1890s.)

Question 6: Is Euclid's reasoning sound? Suppose we try to verify his theorem on a plane of points with rational coordinates. Let A be the origin $(0, 0)$ and $B = (0, 1)$. What happens?

Euclid's proof contains justifications for all its statements except one: Let C be a point of intersection of the circles. It happens that no postulate or theorem deducible from Euclid's postulates ensures that the circles intersect. An *axiom of continuity* fixes this gap. This axiom was first proposed by Dedekind.

Axiom **Axiom of Continuity:**

D-1: Let a line ℓ be partitioned into two nonempty subsets of points, S and T, in such a way that no point of S is between two points of T and no point of T is between two points of S. Then there exists a unique point O on ℓ with the following property: For any points A and B on ℓ distinct from O, O is between A and B if and only if A is in S and B is in T, or A is in T and B is in S.

Prior to Dedekind, most mathematicians had based their understanding of the continuity of a line on the fact that between any two points there exists another point

(Axiom B-1). But Dedekind realized that between any two rational numbers there exists a rational number, and yet the rational numbers do not form a continuum. The real numbers, on the other hand, do form a continuum and can be put into one-to-one correspondence with the points of a line.

Dedekind recognized that the continuity of a line can be based on a line separation property, i.e., the fact that a point can partition a line into two sets of points with the property that each point of the line belongs to one and only one set, and all the points of one set are to the left of all the points in the other set. Dedekind's axiom can be thought of as the converse to this line separation property: There is a unique point that creates that partition, separating the points of the line to the left and right of it.

The Axiom of Continuity guarantees that two circles have a point of intersection, but the argument is too long for us to give the details here. Instead, we outline a proof that is detailed in Heath's commentary on Euclid's *Elements*.

1. Call a ray an **interior ray** of an angle if its endpoint is the vertex of the angle and if it is between the sides of the angle. Consider the set I of all interior rays of an angle. Using the Axiom of Continuity, it can be proved that if a ray R in set I partitions I into itself and two subsets S and T so that all rays in S are on one side of the ray R, and all rays in T are on the other side of R, then R is either at the edge of S or at the edge of T.
2. If an angle is a central angle of a circle, the same partition of its interior rays also partitions the intersections of these rays with the circle itself. Consequently, the points of a minor arc can be partitioned just as the rays in (1) can be partitioned.
3. From (2) it can be proved that if a line has one point in the interior and one point in the exterior of a circle, then it has two points in common with the circle.
4. From (1), (2), and (3) it can be proved that if a circle A contains one point X that is in the interior of circle B, and if circle A contains a second point Y that is in the exterior of circle B, then circles A and B must intersect in two points. (See Problem 5.)

11.1.3 Problems

1. Draw a diagram illustrating each statement.
   a. Axiom C-1
   b. Axiom C-4
   c. definition of "is greater than" for angles

2. Using only the language developed in Sections 11.1.2 and this section, define each term.
   a. complementary angles
   b. obtuse angle

3. How does the definition of triangle in this section differ from Euclid's definition shown in Section 7.1.1?

4. Each of the following terms is defined in this section. Reword that definition using the concepts of distance and/or angle measure.
   a. circle
   b. supplementary angles
   c. right angle

5. a. Using the concept of betweenness and without using the concept of distance, define the interior of a circle and the exterior of a circle.
   b. Use the axioms we have assembled so far to explain why there must exist two points X and Y that are in the interior of a circle with center O such that O is between X and Y.

6. Use Theorem 11.2 (the Postulate of Pasch) to prove that no line can contain points on all three sides of a triangle without containing a vertex of the triangle.

7. Use the triangle congruence axiom (C-7) to prove that the base angles of an isosceles triangle are equal. This is a famous theorem called *pons asinorum* or *Bridge of Asses*. (It receives its name from the diagram Euclid used in its proof.)

8. Using the betweenness axioms and the definitions of segment and ray, prove that the intersection of $\overrightarrow{AB}$ and $\overrightarrow{BA}$ is $\overline{AB}$.

9. A student argued that the Plane Separation Axiom B-5 can be converted into the line separation property merely by replacing the word "plane" by the word "line" and "line" by "point". Is the student correct? Why or why not?

10. Prove the following theorem: Suppose $\overline{AB} \cong \overline{CD}$. For any point E between A and B, there exists a unique point F between C and D such that $\overline{AE}$ is congruent to $\overline{CF}$.

11. Prove the following theorems using the segment congruence axioms and the definition of > (greater than) for segments.

a. Exactly one of the following conditions holds: $\overline{AB} > \overline{CD}$, $\overline{AB} \cong \overline{CD}$, or $\overline{CD} > \overline{AB}$.

b. If $\overline{AB} > \overline{CD}$ and $\overline{CD} > \overline{EF}$, then $\overline{AB} > \overline{EF}$.

With Theorem 11.5, these theorems provide for the ordering of segments.

12. Prove Theorem 11.6.

13. Prove the following theorems:

a. If P is in the interior of ΔABC, the line $\overleftrightarrow{AP}$ intersects the line segment $\overline{BC}$.

b. If P is in the interior of ΔABC, any line through P intersects two of the sides of the triangle.

14. Prove the following theorem: Suppose $\angle ABC \cong \angle DEF$. For any point X between A and C, there exists a unique point Y between D and F such that $\angle ABX \cong \angle DEY$. (*Hint*: Use Problem 10.)

15. Use Theorem 11.7 to prove the following theorems:

a. Vertical angles are congruent to each other.

b. An angle congruent to a right angle is a right angle.

16. Use the definition of > for angles to create and prove two theorems on the ordering of angles analogous to those for the ordering of segments in Problem 11.

ANSWERS TO QUESTIONS

1. Axiom C-2 replaces Euclid's first and fourth common notions. Axiom C-3 replaces his second common notion.

2. Axiom C-1 provides for them. We need three points.

3. Axiom B-3.

4. No. Let $\angle AOB$ be a right angle. Let $\angle COD$ be a right angle with $\overrightarrow{OC}$ not in the interior of $\angle AOB$, C and B on the same side of $\overleftrightarrow{OA}$, and C and D on opposite sides of $\overleftrightarrow{OA}$. Then $\overrightarrow{OB}$ is between $\overrightarrow{OA}$ and $\overrightarrow{OC}$, and $\overrightarrow{OC}$ is between rays $\overrightarrow{OB}$ and $\overrightarrow{OD}$, but $\overrightarrow{OB}$ is not between rays $\overrightarrow{OA}$ and $\overrightarrow{OD}$.

5. The interior of $\angle AOB$ is the set of points X such that $\overrightarrow{OX}$ is between $\overrightarrow{OA}$ and $\overrightarrow{OB}$. It can also be described as the intersection of two half planes: the half-plane of B bounded by $\overleftrightarrow{OA}$ and the half-plane of A bounded by $\overleftrightarrow{OB}$.

6. First construct a circle with center $(0, 0)$ and radius 1, then another circle using point $(0, 1)$ as the center and radius 1. These two circles do not intersect in a point with rational coordinates, so Euclid's proof breaks down on the given plane.

11.1.4 Geometry without the Parallel Postulate

All the axioms we have stated so far are valid propositions in both Euclidean and hyperbolic geometry. In the next section we add a final postulate that is not true in hyperbolic geometry and that allows us to define Euclidean geometry completely.

This postulate is Euclid's parallel postulate. However, we first examine some of the results of Euclidean (and hyperbolic) geometry that can be achieved without the parallel postulate. There are good reasons to do this. Euclid proved Propositions 1 through 28 in the first book of the *Elements* without using his parallel postulate. (See the list in Section 7.1.1.) By examining the body of theorems that do not depend on a parallel postulate we are better able to understand the role that the postulate plays in Euclidean geometry.

In particular we want to look at some important concepts and theorems of Euclidean geometry, see how they are related to each other, and determine if they require the parallel postulate. We first prove a familiar property of isosceles triangles. We call a triangle **isosceles** if and only if it has at least two congruent sides. The third side is called the **base** and the angles that include it are the **base angles** of the triangle. This proof of Theorem 11.8 is due to Pappus.

Theorem 11.8 Base angles of an isosceles triangle are congruent.

Proof: Let ΔABC be isosceles with $\overline{AB} \cong \overline{AC}$. We wish to show that $\angle ABC \cong \angle ACB$. Consider ΔABC and ΔACB. $\overline{AC} \cong \overline{AB}$ by Axiom C-2. Also, by Axiom C-5, $\angle CAB \cong \angle BAC$. Consequently, $\angle ABC \cong \angle ACB$ by the SAS Congruence Axiom C-7. ∎

Theorem 11.9 **(SSS Congruence):** If $\overline{AB} \cong \overline{DE}$, $\overline{BC} \cong \overline{EF}$, and $\overline{AC} \cong \overline{DF}$, then $\Delta ABC \cong \Delta DEF$.

Proof: By Axiom C-4, in the half plane of line $\overleftrightarrow{AB}$ that does not contain C there exists a ray $\overrightarrow{AG}$ such that $\angle BAG \cong \angle FDE$. By Axiom C-1, there is a point C' on $\overrightarrow{AG}$ such that $\overline{AC'} \cong \overline{DF}$, which by Axiom C-5 and the given $\overline{AC} \cong \overline{DF}$ implies $\overline{AC'} \cong \overline{AC}$. By Axiom C-7 (SAS Congruence), $\overline{BC'} \cong \overline{EF}$, which, with Axiom C-5 and the given $\overline{BC} \cong \overline{EF}$, implies $\overline{BC'} \cong \overline{BC}$.

By Axiom B-5, $\overline{CC'}$ intersects $\overleftrightarrow{AB}$ at a point P. There are now five cases.

Case 1: P is between A and B. Then $\Delta CAC'$ and $\Delta CBC'$ are isosceles. Consequently, by Theorem 11.8, $\angle ACC' \cong \angle AC'C$ and $\angle BCC' \cong \angle BC'C$. By Axiom C-6, $\angle ACB \cong \angle AC'B$. So $\Delta ABC \cong \Delta ABC'$ by SAS Congruence. But $\Delta ABC' \cong \Delta DEF$, and since triangle congruence is an equivalence relation, $\Delta ABC \cong \Delta DEF$, which was to be proved.

Case 2: A is between P and B. The argument for this case is like case (1), except that we use Theorem 11.6 instead of Axiom C-6.

Case 3: B is between P and A. The argument is just like case (2), with A and B changing places.

Case 4: $P = A$. Then C', P, and C are collinear. Thus $\Delta BCC'$ is isosceles. So, by Theorem 11.8, $\angle BCC' \cong \angle BC'C$. So $\Delta ABC \cong \Delta ABC'$ by SAS Congruence. The rest is identical to case (1).

Case 5: $P = B$. This argument is just like case (4), with A and B changing places. ∎

Theorem 11.10 **(AAS Congruence):** If $\overline{AB} \cong \overline{DE}$, $\angle ACB \cong \angle DFE$, and $\angle BAC \cong \angle EDF$, then $\Delta ABC \cong \Delta DEF$.

You are asked to prove this theorem in Problem 1.

We have not yet defined some simple ideas, among them *equidistance* and the *midpoint* of a segment.

Definitions A point P is **equidistant** from A and B if and only if $\overline{PA} \cong \overline{PB}$. M is the **midpoint** of $\overline{AB}$ if and only if M is on $\overleftrightarrow{AB}$ and equidistant from A and B.

Question 1: What is the significance of the word *the* in the definition of *midpoint*?

Although from Axiom B-1 we know that, given two different points A and B, there exists a point between them, we do not know automatically that there is one and only one point equidistant from them on the segment. Some geometers make the existence and uniqueness of a midpoint an axiom. Others deduce these properties from assumed metric properties of segments (when we postulate that a line can be coordinatized with real numbers, we have unique midpoints). Regardless of method, proofs of the existence and uniqueness of a midpoint often reveal interesting connections between it and other geometric figures.

You may have seen some proofs about unique midpoints based on the Euclidean parallel postulate or a postulate equivalent to it. But we can construct a proof without it, on the basis of axioms we have collected so far, with the addition of some definitions and theorems.

Theorem 11.11 Every segment has exactly one midpoint.

Proof: Let $\overline{AB}$ be a segment. Using Euclid's Theorem 1 and the Axiom of Continuity, construct equilateral triangles ABC and ABC' with C and C' on opposite sides of $\overleftrightarrow{AB}$. Thus $\overline{AC} \cong \overline{AB} \cong \overline{BC} \cong \overline{BC'} \cong \overline{AC'}$. Let M be the intersection of $\overline{CC'}$ and $\overleftrightarrow{AB}$. We want to show that M is a midpoint of $\overline{AB}$. $M \neq A$, for if $M = A$ then by the same argument starting with $\overline{BA}$ (which equals $\overline{AB}$), $M = B$, and so $A = B$, which is impossible since $\overline{AB}$ is a segment. So A and B do not belong to $\overline{CC'}$, and triangles $AC'C$ and $BC'C$ exist. $\angle AC'C \cong \angle BC'C$ by SSS Congruence. So $\angle AC'M \cong$ $\cong \angle BC'M$ (since M is on $\overline{CC'}$). Then $\overline{AM} \cong \overline{BM}$ by SAS Congruence. So M is a midpoint of $\overline{AB}$. Uniqueness is left to you as Problem 2. ⌟

Question 2: Draw a diagram to illustrate this proof.

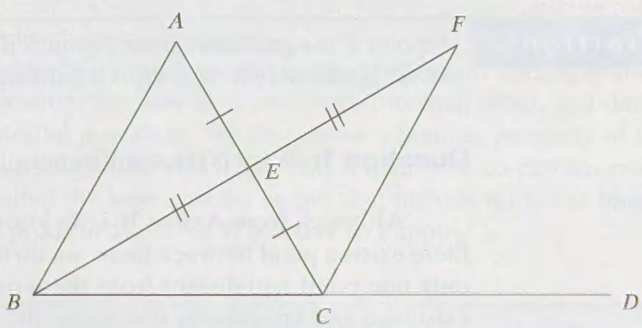
From the proof of Theorem 11.11, we see that the existence of midpoints is related to triangle congruence, incidence, betweenness, and separation properties of the plane. The parallel postulate is not required. To be certain of this, we would, however, need to examine the theorems (such as SSS and AAS) used, as well as, for example, whether anything besides continuity is required in the first proof to construct points C and C'.

It is often the case that proof reveals some surprising relationships between the concepts of geometry. For example, it turns out that the existence and uniqueness of a midpoint plays an important role in the proof of one of the famous theorems of Euclidean geometry—the exterior angle theorem.

Definition In ΔABC, if point D is on $\overleftrightarrow{BC}$ such that C is between B and D, then $\angle ACD$ is called an **exterior angle of ΔABC**.

Theorem 11.12 **(Exterior Angle Theorem):** In any triangle, an exterior angle is greater than either of its nonadjacent interior angles.

Figure 7



Proof: Suppose ΔABC has exterior angle ACD (Figure 7). Let E be the midpoint of side $\overline{AC}$. Extend $\overline{BE}$ to F so that $\overline{BE} \cong \overline{EF}$. The vertical angles AEB and CEF are congruent, and $\overline{AE} \cong \overline{EC}$, so $\angle ACF \cong \angle CAB$ by SAS Congruence (Axiom C-7). F is on the opposite side of $\overleftrightarrow{AC}$ from B, so it is on the same side of $\overleftrightarrow{AC}$ as D (Plane Separation Axiom). Also, since E is between B and F, and since $\overline{FB}$ intersects $\overleftrightarrow{CD}$

at B, F is on the same side of $\overleftrightarrow{CD}$ as E. Thus F is in the interior of $\angle ACD$. Consequently, $\angle ACD$ is greater than $\angle ACF$, so $\angle ACD > \angle CAB$. This same argument could be repeated starting with the midpoint of side $\overline{BC}$, so the theorem is proved. ∎

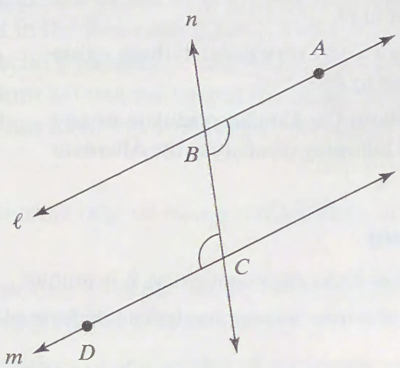
The exterior angle theorem is very powerful. It enables us to prove, for example, that given a line and a point not on it, there is a unique line on that point perpendicular to that line (see Problem 5).

Now we introduce parallelism. We adopt here Euclid's definition of *parallel*: **Parallel lines** are lines with no points in common. A sufficient condition for parallelism is found in an important theorem, which follows from the Exterior Angle Theorem. Before we explore this theorem, we again need some definitions:

Definitions Let lines ℓ, m, and n be distinct. Let B be on ℓ and n, and let C be on m and n such that B is distinct from C. Choose A on ℓ and D on m such that A and D are on opposite sides of n. Then $\angle ABC$ and $\angle BCD$ are a pair of **alternate interior angles** formed by the line n, which we call a **transversal**, and the lines ℓ and m.

Theorem 11.13 **(Alternate Interior Angle Theorem):** If two lines are intersected by a transversal so that alternate interior angles are congruent, then the lines are parallel.

Figure 8



Proof: The proof is indirect. Refer to Figure 8. Assume that the alternate interior angles ABC and DCB are congruent but $\overleftrightarrow{AB}$ and $\overleftrightarrow{CD}$ intersect. Then find a contradiction to one of the theorems we have proved. The rest is left to you. ∎

Since Theorem 11.12 was used in the proof of Theorem 11.13, can we conclude that the exterior angle theorem is necessary to prove the alternate interior angle theorem, and therefore that the concept of midpoint is also necessary for its proof? No. Consider the following proof of the alternate interior angle theorem:

Proof: Let ℓ, m, n and points A, B, C, D be as in the preceding definition of alternate interior angles and $\angle ABC \cong \angle DCB$. Assume that ℓ and m meet at a point E, and that E is on the same side of n as D and the opposite side of n as A. There is a point F on $\overrightarrow{BA}$ such that $\overline{BF} \cong \overline{CE}$ (Axiom C-1). $\overline{CB}$ is congruent to itself (Axiom C-2) so in triangles BCE and CBF, $\angle EBC \cong \angle FCB$ (Axiom C-7). Since $\angle EBC$ is the supplement of $\angle FBC$, $\angle FCB$ is the supplement of $\angle ECB$ (Axiom C-4, Theorem 11.7). Then F is on m. Since both E and F are on ℓ and m, $\ell = m$ (Axiom I-1). This contradicts the given that ℓ and m are distinct lines. Therefore, ℓ is parallel to m. ∎

Question 3: Can we conclude from the Alternate Interior Angle Theorem that parallel lines exist?

We did not use the concept of midpoint in the previous proof. In fact, it turns out that we can use the alternate interior angle theorem to construct yet another proof of the existence of a midpoint, again based on the axioms we have collected so far.[1]

This discussion is intended to help you understand the power of proof in exposing relationships between geometric concepts. It also serves as a caution to you not to assume that simply because a result is used in the proof of a theorem, it is necessary for all proofs of that theorem. It might simply be sufficient. For example, there are proofs of the exterior angle theorem based on Euclid's parallel postulate, but, as we have shown, the parallel postulate is not necessary for the proof.

11.1.4 Problems

1. Prove Theorem 11.10 (AAS Congruence).

2. Prove the uniqueness of the midpoint of a line segment. (*Hint*: Assume there are two midpoints and derive a contradiction using AAS.)

3. Prove ASA Congruence using SAS Congruence.

4. Complete the proof of Theorem 11.13.

5. If $\angle AOB$ is a right angle, then lines $\overleftrightarrow{OA}$ and $\overleftrightarrow{OB}$ are **perpendicular** to each other at O.
   a. Prove that for every line ℓ and every point P there exists a line on P perpendicular to ℓ.
   b. Use part **a** to prove (without the parallel postulate or any equivalent statement) the following corollary to the Alternate Interior Angle theorem: If ℓ is any line and P is any point not on ℓ, there exists at least one line m through P parallel to ℓ.
   c. Prove that the perpendicular in part **a** is unique.

6. Use the following theorem (which can also be proved without the parallel postulate) to give an alternate proof of the corollary to the Alternate Interior Angle Theorem stated in Problem 5**b**.

7. Prove that the alternate interior theorem, in the presence of the incidence, betweenness, and congruence axioms, is sufficient to show that parallel lines exist. You may assume that all right angles are congruent. (*Hint*: Use perpendicular lines.)

ANSWERS TO QUESTIONS

1. The word *the* signifies that if the midpoint exists, it is unique.

3. Yes, because congruent alternate interior angles can be formed at points B and C, the theorem is sufficient to prove the existence of parallel lines.

11.1.5 Euclid's Fifth Postulate

Although many theorems can be proved from the axioms we have so far stated, which fill in the gaps in the first four postulates of Euclid, not all the theorems of Euclidean geometry can be deduced from them. Like Euclid, when he was using only the first four of his postulates, we are one postulate short of those needed to prove all the theorems found in Euclid's *Elements*. That is, we need one more postulate in order to have a full set of postulates for Euclidean geometry. Here again is Euclid's fifth postulate.

Axiom **Euclid's fifth postulate (parallel postulate):**

P-E: If a straight line falling on two straight lines makes the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which are the angles less than the two right angles.

[1] See Marvin J. Greenberg, *Euclidean and Non-Euclidean Geometries* (Third Edition). San Francisco: W. H. Freeman, 1993, p. 137.

Question: Draw the figure described in the fifth postulate.

In the English translation (by Thomas Heath) that we have adapted for use here, the fifth postulate uses more words than Euclid's first four postulates put together. This also is the case in the original Greek, so it should not surprise you that, from the time of Euclid, the length and complexity of the fifth postulate led mathematicians to try to prove it from the other postulates. Examination of Euclid's first 28 propositions suggests that he also may have tried to prove his fifth postulate from the other four.

It was not until the 19th century that such attempts were revealed to be futile because models of geometry were described that satisfied the first four postulates but not the fifth. Still, much was learned from efforts to prove the fifth postulate, including the discovery of non-Euclidean geometries and the proofs of many statements equivalent to the parallel postulate of Euclid.

No original version of Euclid's *Elements* survives, and our knowledge of Euclid comes mainly from two detailed commentaries written by Greeks centuries later, Pappus (c. 320 A.D.) and Proclus (410–485 A.D.). Proclus was one of the first to advocate that Euclid's fifth postulate be a theorem, proved from a postulate he suggested:

1. There is only one parallel to a given line through a point not on that line.

This has become known as Playfair's version of the parallel postulate, after John Playfair (1748–1819), who suggested its use in 1795. We noted in Section 11.1.2 that Playfair's postulate is violated in the Beltrami-Cayley-Klein model of hyperbolic geometry. Thus once we include Euclid's parallel postulate (or any statement, such as Playfair's, equivalent to it), our axiom set can no longer describe hyperbolic geometry.

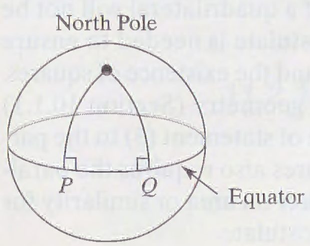
Other statements that have been proved to be equivalent to the fifth postulate include the following:

2. Any line that intersects one of two parallel lines intersects the other. (Proclus)
3. Given a triangle, another triangle can be constructed that is similar and not congruent to it. (John Wallis, 17th century)
4. The sum of the measures of the angles of a triangle equals 180°.

Euclid's fifth postulate does not mention the word "parallel", yet it is sometimes called Euclid's "parallel postulate" because it has implications for the behavior of parallel lines. Since Euclid had defined "angle" in terms of "inclination", he might have defined "parallel lines" as lines with the same inclination. But he never uses this word again and would have found it difficult to prove things about parallel lines had he adopted such a definition. He also could have defined parallel lines as lines that "go in the same direction", perhaps like chariot tracks. But he may have realized that the Earth is nearly a sphere (Greek mathematicians *did* know Earth was round) and that two lines drawn in the directions of due north and south would intersect at the poles. He might have defined parallel lines as lines that are equidistant, but this definition requires knowing that the set of points at a fixed distance from a line on one side of it is in fact a line. Instead, Euclid defines lines to be *parallel* if they are in the same plane and do not meet. Then he assumes the fifth postulate, which provides a sufficient condition for lines not to be parallel!

Euclid's fifth postulate shows once again that Euclidean geometry does not apply on the surface of the sphere. Recall the triangle NPQ from Section 10.1.1 (Figure 9) whose sides are the great circles of the sphere. Such a triangle includes two right angles, and so violates Euclid's fifth postulate.

Figure 9



The connection between the fifth postulate and parallelism comes through the logic of the **Principle of the Contrapositive**: If p and q are statements, then $p \Rightarrow q$ has the same truth value as its contrapositive not $q \Rightarrow$ not p.

Euclid's fifth postulate is of the form $p \Rightarrow q$, where

p is "A straight line falling on two straight lines makes the interior angles on the same side less than two right angles."

and q is "The two straight lines, if produced indefinitely, meet on that side on which are the angles less than the two right angles."

So the contrapositive of the fifth postulate is

not $q \Rightarrow$ not p: If two lines do not intersect on a side of a third line, then the interior angles on that side are either equal to or greater than two right angles.

But if the interior angles on one side are greater than two right angles, the interior angles on the other side of the line are less than two right angles, and (due to the postulate) the lines will meet there. So we can conclude: If two parallel lines are cut by a transversal, then the interior angles on the same side of the transversal equal two right angles. That is, (in modern language) they are supplementary.

This argument shows that the parallel postulate is also equivalent to the following statements:

5. If two parallel lines are cut by a transversal, then corresponding angles are congruent.
6. If two parallel lines are cut by a transversal, then alternate interior angles are congruent.

The converses of statements (5) and (6) are not equivalent to the parallel postulate. Theorem 11.13 shows that (6) can be deduced from our axioms without the parallel postulate.

Theorems requiring the Parallel Postulate

We can prove many theorems of Euclidean geometry without the parallel postulate, such as, for example, that two lines perpendicular to the same line are parallel to each other. However, the following theorem, which appears to be quite similar, requires it.

Theorem 11.14 Two distinct lines parallel to the same line are parallel to each other.

Proof: Given lines ℓ, m, n, such that ℓ is parallel to m and n is parallel to m, and lines ℓ and n are distinct. We prove that ℓ is parallel to n. Suppose not. Let A belong to ℓ and n. Since ℓ is distinct from n, there exist two distinct lines, ℓ and n, each on A and each parallel to m. But this contradicts the Euclidean parallel postulate (Playfair's version). Therefore, ℓ is parallel to n. ∎

You may be surprised at other theorems that require the parallel postulate. Because of their equivalence to the parallel postulate, statements (1) to (6) require the parallel postulate or some equivalent assumption in order to be proved. Because of statement (4), the sum of the measures of the angles of a quadrilateral will not be 360° without the parallel postulate. Thus the parallel postulate is needed to ensure the existence of rectangles (figures with four right angles) and the existence of squares. Consequently, the basic definition of area in Euclidean geometry (Section 10.1.1) requires the parallel postulate. Because of the equivalence of statement (3) to the parallel postulate, the existence of noncongruent similar figures also requires the parallel postulate. The Pythagorean Theorem, which relies either on area or similarity for its proof (Section 8.3.1), thus also relies on the parallel postulate.

With the parallel postulate, not only do we get all the above theorems, but we have an axiom set that defines Euclidean plane geometry. That is, using the 18 axioms I1–I4, B1–B5, C1–C7, D-1, and P-E, and the definitions we have given, all the theorems associated with Euclidean plane geometry can be deduced.

11.1.5 Problems

1. Consider this statement: If a line is perpendicular to one of two parallel lines, then it is perpendicular to the other. Prove that it is implied by Euclid's parallel postulate.

2. Prove that statement (5) in this section is equivalent to Euclid's parallel postulate.

3. Prove that the suggested alternates (1) and (2) to the parallel postulate mentioned on the first page of this section imply each other.

4. Prove that Playfair's parallel postulate (1) implies statement (6) in this lesson. (*Hint*: Use Theorem 11.10.)

5. Determine if the parallel postulate holds in the following geometries mentioned in Section 11.1.1. Support your answers.

a. Fano's geometry

b. Young's geometry (see Problem 9 of Section 11.1.1)

c. three-point geometry (see Problem 1 of Section 11.1.1)

d. four-point geometry (see Problem 11 of Section 11.1.1)

6. Consider the geometry consisting of 13 points $\{A, B, C, D, E, F, G, H, I, J, K, L, M\}$ and 26 lines: $\{ABC, BDF, CDG, DHI, EFK, FGM, GIK, HJL, ADE, BEI, CEJ, DJK, EGL, FIJ, AFH, BGH, CFL, DLM, EHM, AGJ, BJM, CHK, AIL, BKL, CIM, AKM\}$. Here ABC means the line $\{A, B, C\}$, etc.

a. Does the Euclidean parallel postulate hold?

b. Prove or disprove that this geometry satisfies the incidence axioms.

c. Formulate a theorem in the above geometry.

d. Does the postulate of Pasch (Theorem 11.2) hold in this geometry?

7. Answer the same questions as Problem 6 for the geometry consisting of 13 points $\{A, B, C, D, E, F, G, H, I, J, K, L, M\}$ and 13 lines $\{ABCD, AEFG, AHIJ, AKLM, BEHK, BFIL, BGJM, CEIM, CFJK, CGHL, DEJL, DFHM, DGIK\}$. Here $ABCD$ means the line with four points $\{A, B, C, D\}$.

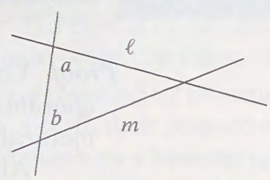
8. a. Prove that the perpendicular bisectors of the three sides of a triangle have a point in common.

b. Identify the place(s) in your proof where the parallel postulate is used.

ANSWER TO QUESTION

Figure 10 is one such figure.

Figure 10



If $a + b < 180$, ℓ and m intersect.

Unit 11.2 The Cartesian Model for Euclidean Geometry

In this unit, we explore an approach to Euclidean geometry that is different from the synthetic approach exemplified by the axioms I1–I4, B1–B5, C1–C7, D-1, and P-E of Unit 11.1. This approach is analytic, and by examining it, we connect the lines that are studied in algebra with the lines in Euclidean geometry. It also justifies the analytic proofs that we have in this book for deducing some of the theorems of Euclidean geometry.

In this analytic approach to Euclidean geometry, we are led to ask questions about our axiom system for Euclidean geometry. Does our axiom system apply only to Euclidean geometry, or, like the incidence axioms, can it apply to different geometries?

11.2.1 The Cartesian coordinate system

Recall from Section 8.1.1 that the **Cartesian plane** is the set $\mathbf{R}^2$ of all ordered pairs of real numbers. Accordingly, a Cartesian **point** is an ordered pair of real numbers. A Cartesian **line** is the set of points (x, y) that satisfy an equation equivalent to a linear equation in the standard form $Ax + By + C = 0$, where A and B are constant real numbers that are not both zero. If $A = 0$, then the equation represents a **horizontal line**; if $B = 0$, a **vertical line**. While you can think of drawing the usual picture of such a system with the x-axis as horizontal and y-axis as vertical, all of the work that we do in this system in this section can be done with algebra alone, dependent not on the pictures but on the properties of real numbers.

Note, that the standard form $Ax + By + C = 0$ is not unique, since all multiples $kAx + kBy + kC = 0$, where k is a nonzero constant, represent the same line. However, a line is uniquely determined by its slope and y-intercept. So if the line is not vertical, then $B \neq 0$ and we can convert the standard form of the equation to the familiar slope-intercept form of an equation: $y = mx + b$, where m is the slope of the line and b is the y-intercept, where $m = -\frac{A}{B}$ and $b = -\frac{C}{B}$. If $B = 0$, we say the line has *no slope*, and the equation is of the form $x = h$, where $h = -\frac{C}{A}$. The equation $kAx + kBy + kC = 0$, when converted to slope-intercept form, yields the same slope and y-intercept as the original equation. Therefore, all such multiples can be treated as essentially the same equation.

We wish to demonstrate that the Cartesian plane with these points and lines is a *model* of Euclidean geometry, that is, an example or interpretation or manifestation of the axiom system of Euclid's geometry. To do this, we must show that each axiom in Euclidean geometry is satisfied in the model. Thus the axioms in our earlier development of Euclidean geometry become theorems in the model. Our first theorems are the four incidence axioms.

Theorem 11.15 **(Cartesian I-1):** There exist at least three distinct points.

Proof: We have an infinity of trios of Cartesian points from which to select. One trio is $(1, 1)$, $(0, -4)$, and $\left(\frac{3}{8}, 7\right)$. ⌟

Theorem 11.16 **(Cartesian I-2):** For each two distinct points, there exists a unique line on them.

Proof: Let (x_1, y_1) and (x_2, y_2) be the two distinct points. Then, algebraically, the coordinates of these points must satisfy the equation of any line on them. Geometrically, the line passes through the two points.

An equation for the unique line on these two points is found by using the familiar two-point formula

$$y - y_1 = \frac{y_1 - y_2}{x_1 - x_2}(x - x_1),$$

which we can use when x_1 does not equal x_2. How do we know this is the *only* equation that describes the line on these two points? We convert our equation into standard form.

$$-(y_1 - y_2)x + (x_1 - x_2)y + (y_1x_2 - x_1y_2) = 0$$

Considering all multiples as representing the same line, we see that A, B, and C are uniquely determined by the coordinates of the given points. That is,

$$A = -(y_1 - y_2), \quad B = (x_1 - x_2), \quad \text{and} \quad C = (y_1x_2 - x_1y_2).$$ ⌟

Question 1: Prove Cartesian I-2 for the case $x_1 = x_2$.

Theorem 11.17 **(Cartesian I-3):** For every line there exist at least two distinct points on it.

Proof: Let the line have equation $Ax + By + C = 0$. First assume A is zero, but B is not zero. Two possible points are $\left(1, -\frac{C}{B}\right)$ and $\left(2, -\frac{C}{B}\right)$. If B is zero, and A is not zero, our points can be $\left(-\frac{C}{A}, 13\right)$ and $\left(-\frac{C}{A}, 0\right)$, among others. If neither A nor B is zero, we could choose $\left(\frac{C}{A}, -\frac{2C}{B}\right)$ and $\left(-\frac{2C}{A}, \frac{C}{B}\right)$. ⌟

Theorem 11.18 **(Cartesian I-4):** Not all points are on the same line.

Proof: Again, we have an infinity of points to select in order to demonstrate this axiom. The points we chose in the proof of Cartesian I-3 will suffice. ⌟

Thus the Cartesian plane satisfies the incidence axioms of Euclidean geometry. What about the Euclidean parallel postulate? Let's check on this. We'll use the Playfair version: On a given point not on a given line, exactly one line can be drawn parallel to a given line. We must first specify what it means in our model for two lines to be parallel. We define parallel lines here as lines that do not intersect, so two lines in our model are parallel if there is no ordered pair that satisfies their respective equations.

Before we demonstrate the Playfair axiom, we first prove three lemmas.

Lemma 1 Every vertical line intersects every nonvertical line.

Proof: Any vertical line can be represented by the equation $x = A$, and any nonvertical line can be represented by the equation $y = mx + b$. The ordered pair that satisfies both these equations is $(A, mA + b)$. ⌟

Lemma 2 Two lines are parallel if and only if (1) both are vertical lines; (2) neither is vertical and they both have the same slope.

Proof: Since Lemma 2 is an "if and only if" statement, we must prove both directions.

($\Leftarrow$) (1) If both distinct lines are vertical or (2) if neither of two distinct lines is vertical and both have the same slope, we must prove that the two lines are parallel.

(1) Let two distinct vertical lines be represented by the equations $x = A$ and $x = B$, where $A \neq B$. Since $A \neq B$, their respective solutions sets, $\{(A, y)\}$ and $\{(B, y)\}$ for all real numbers y have no common solution. Therefore, the lines are parallel.

(2) Let two distinct nonvertical lines with the same slope be represented by the equations $y = mx + b_1$ and $y = mx + b_2$, where $b_1 \neq b_2$. Solving these equations simultaneously produces a contradiction: $b_1 = b_2$, and so there is no common solution. The lines are parallel.

($\Rightarrow$) If two lines are parallel, we must prove that either (1) both lines are vertical or (2) if neither line is vertical, both have the same slope.

Let two lines be parallel. Then no ordered pair is a common solution to their respective equations. By Lemma 1, then either both lines are vertical or neither line is vertical. If both lines are vertical, then (1) holds. Therefore, it remains only to prove that if neither line is vertical, both must have the same slope in order to be parallel.

We proceed using an indirect proof. Assume two nonvertical lines have different slopes yet are parallel. Let the first line be represented by the equation $y = m_1x + b_1$. Let the second line be represented by the equation $y = m_2x + b_2$, where m_1 is not equal to m_2. Solving these equations simultaneously, we obtain

$$m_1x + b_1 = m_2x + b_2,$$

from which

$$x = \frac{b_2 - b_1}{m_1 - m_2}$$

and

$$y = m_1\left(\frac{b_2 - b_1}{m_1 - m_2}\right) + b_1.$$

So $(x, y) = \left(\frac{b_2-b_1}{m_1-m_2}, m_1\left(\frac{b_2-b_1}{m_1-m_2}\right) + b_1\right)$. These values for x and y satisfy the equations of both lines ($y = m_1x + b_1$ and $y = m_2x + b_2$), indicating that the lines intersect in the point represented by that ordered pair. This contradicts our hypothesis that the two lines are parallel, and so our assumption that they had different slopes must be false. ┘

Lemma 3 Given a point $P = (x_1, y_1)$ and a real number m, there is exactly one line that is on P and has slope m.

Proof: Lines with slope m can be represented by equations of the form $y = mx + b$. If such a line contains P, then the coordinates of P must satisfy its equation. Substituting x_1 and y_1 for x and y, respectively, we obtain $b = y_1 - mx_1$. Therefore, the unique line on P with slope m is $y = mx + (y_1 - mx_1)$. ┘

The lemmas we proved have established certain properties of vertical and nonvertical lines in the Cartesian plane. We are now ready to prove that the Playfair axiom holds in the Cartesian plane. In the proof below, we partition the lines of the Cartesian plane into vertical and nonvertical, treating each as a separate case.

Theorem 11.19 **(Cartesian Parallel Postulate):** On a given point not on a given line, exactly one straight line can be drawn parallel to a given straight line.

Proof: Given a line and a point $P = (x_1, y_1)$ not on it.

Case 1. The given line is vertical. Let this line be represented by the equation $x = A$. Then the unique vertical line on P is given by the equation $x = x_1$. Since both lines are vertical, by Lemma 2, they are parallel, and by Lemma 1 there is no nonvertical line parallel to our given line. Thus the Playfair axiom is verified for this case.

Case 2. The given line is not vertical. Let this line be represented by the equation $y = mx + b$. Then by Lemma 2, the only line on P parallel to the given line is a line on P with slope m. Using Lemma 3, this unique line is given by the equation $y = mx + (y_1 - mx_1)$. Thus the Playfair axiom is also verified for this case. ┘

We have thus established that the incidence axioms and the parallel postulate (Playfair version) are satisfied in the Cartesian plane. To complete the proof that the Cartesian plane is a model of Euclidean plane geometry as defined by our axiom system, we would have to show that all the other axioms of betweenness, congruence, and continuity are similarly verified. Some of these proofs are included as problems and others can be found in the literature.[2]

By demonstrating that all of the 18 axioms we gave in Unit 11.1 are true in the Cartesian plane, we allow coordinates to be used for proofs in Euclidean geometry.

11.2.1 Problems

1. For each of the cases in the proof of Cartesian I-3, find two points on the line $Ax + By + C = 0$ other than the two points shown.

2. Give an equation for the line through the two given points.

a. (a, b) and (c, d) b. (x_0, y_0) and $(0, 0)$
c. $(a, 0)$ and $(0, b)$ d. (x_1, y_1) and (x_1, y_2)
e. $(a, -a)$ and $(-b, b)$

3. In the proof of Cartesian I-4, it is asserted that $(1, 1)$ $(0, -4)$, and $\left(\frac{3}{8}, 7\right)$ are not on the same line. Show that this is true with two different explanations.

4. Give an equation for the line parallel to the given line through the given point.

a. $\{(x, y)\colon\ Ax + By + C = 0\}$; $(0, 0)$
b. $\{(x, y)\colon\ x = 1\}$; $(2, 5)$
c. $\{(x, y)\colon\ ax + by = 1\}$; $(c, 0)$
d. $\{(x, y)\colon\ 4y = 2x + 3\}$; $\left(\frac{1}{2}, \frac{1}{3}\right)$

5. Provide an analytic proof of Euclid's Theorem 1 (Section 11.1.3), taking $A = (0, 0)$ and $B = (b, 0)$.

6. Let $A = (x_1, y_1)$ and $C = (x_2, y_2)$, with $A \neq C$. Define B to be between A and C if and only if there exists t such that $0 < t < 1$ and $B = ((1 - t)x_1 + tx_2, (1 - t)y_1 + ty_2)$. From this definition and properties of real numbers, deduce the indicated betweenness axiom from Section 11.1.2.

a. B-1 b. B-2 c. B-3 d. B-4

***7.** Prove that the Postulate of Pasch (Theorem 11.2) is true in the Cartesian plane. (*Hint*: you must use the betweenness axioms and specify an interpretation of segment in terms of distance.)

8. Show that Theorem 11.1 of Section 11.1.1 is true in the Cartesian plane.

9. Consider the system of integers modulo 3. This system consists of the elements 0, 1, 2 with the definitions of multiplication and addition as in Figure 11 (see Unit 6.1).

Figure 11

×	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

+	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

All possible points (x, y) are given by $(0, 0)$, $(0, 1)$, $(0, 2)$, $(1, 0)$, $(1, 1)$, $(1, 2)$, $(2, 0)$, $(2, 1)$, and $(2, 2)$. We can put these nine points into one-to-one correspondence with the nine points of Young's geometry (see Problem 8 of Section 11.1.1). List all possible linear equations. Show that each line consists of exactly three points.

10. Use the axioms of Sections 11.1.1, 11.1.2, and 11.1.3 to construct the Cartesian coordinate system. You may assume certain principles, including the fact that the points on any line can be numbered so that number differences measure distances and the fact that we can assign direction. (*Hint*: Fix an arbitrary point as the origin (indicating what axiom justifies its existence). Choose an arbitrary direction (by convention we would choose from left to right), and an arbitrary unit of length to construct a real number line. Construct a second number line at right angles to the first, justifying its existence. From there, define the coordinates of a point, and deduce the straight line and two of its representations: the two-point form and slope-intercept form.)

ANSWER TO QUESTION

If $x_1 = x_2$, then $y_1 \neq y_2$ because the points are distinct. So (x_1, y_1) and (x_1, y_2) are on the line. The line containing these points has no slope; it is the vertical line $x = x_1$.

[2] See, for example, Edwin E. Moise, *Elementary Geometry from an Advanced Standpoint* (Reading, MA: Addison-Wesley, 1963), Chapter 26.

11.2.2 Verifying the definition of Euclidean geometry: the relationship between a mathematical theory and its models

The Cartesian plane is a model of Euclidean geometry. What does this tell us about Euclidean geometry, in general, and about other models of it? If we prove that certain statements hold in the Cartesian plane, does that mean that these statements are always true in Euclidean geometry? If there are other models of Euclidean geometry, what is their relationship to the Cartesian plane? If a statement is true in the Cartesian plane, will it be true in all other models of Euclidean geometry?

In Sections 11.1.1 to 11.1.5, we defined Euclidean geometry by means of a set of axioms. There are, however, other axiom sets that can define Euclidean geometry. The relationship of such axiom sets to the one we constructed is that each must produce the same body of knowledge (theorems). But how can we determine if another axiom set defines Euclidean geometry? If we accept our axiom system as a defining Euclidean geometry, we must be able to demonstrate that every one of its axioms is true (because it is either taken as an axiom or is provable as a theorem) in any other axiom set claiming to define Euclidean geometry.

Is there any other way to demonstrate that another set of axioms defines Euclidean geometry? Can we use the Cartesian coordinate system? It turns out that the answer is yes, because our axioms for Euclidean geometry enjoy a special property. They are *categorical.*

When an axiom set is **categorical**, it is possible to set up a one-to-one correspondence (isomorphism) between the objects of any two models of that axiom set in such a way that any property that holds for one model will hold for the other. Thus, *every* model of a categorical set of axioms exhibits the same properties.

The property of being categorical, which is applied both to a theory and its axiom sets, was first suggested by John Dewey (1859–1952). Oswald Veblen (1880–1960) introduced it in his systems of axioms for geometry in 1904. Since then, the term and the notion itself have been attributed to Veblen. However, the American mathematician Edward V. Huntington (1874–1952) is sometimes credited with being the first to state it clearly and use it. The first proof of what we today call categoricity, due to Dedekind in 1887, was about axioms for the real numbers.

A model becomes a very powerful tool when we are dealing with a categorical set of axioms. An important theorem in mathematical logic tells us that a statement is true in a mathematical theory if and only if it is true in *every* one of its models. Therefore, when a theory is categorical (where all of its models are essentially the same), we can determine what is provable in that theory on the basis of what we can verify for any single model of it!

Not all axiom systems for Euclidean geometry are categorical. The one we presented is because it includes Axiom D-1 (continuity), which provides sufficient points on a line to be put into 1-1 correspondence with the real numbers. There are other axiom systems for Euclidean geometry that are categorical because of the addition of axioms other than D-1 that allow the establishment of an isomorphism between the geometric line and the real numbers. Since our axiom system for Euclidean geometry is categorical, and we have demonstrated that the Cartesian coordinate system gives a model of this axiom system, we can say that all models of it are essentially the same as its analytic model based on the Cartesian coordinate system. This means that we can use the Cartesian model to make certain statements that we know will be provable from our axioms for Euclidean geometry, and will be valid for every model of it. For example, we can prove the following theorem of Euclidean geometry by verifying that this statement holds in the Cartesian plane.

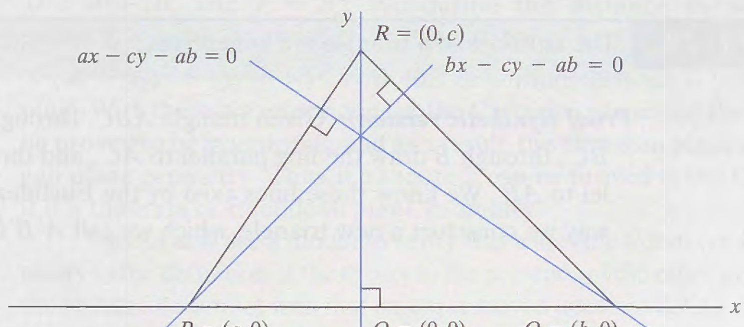
Theorem 11.20 The altitudes of a Euclidean triangle are concurrent.

Proof (analytic version): A coordinate system can be located so that the triangle has vertices $P = (a, 0)$, $Q = (b, 0)$, and $R = (0, c)$, with $abc \neq 0$. Then $\overleftrightarrow{QR}$ has slope $-\frac{c}{b}$, and $\overleftrightarrow{PR}$ has slope $-\frac{c}{a}$. The altitude of ΔPQR from P is perpendicular to $\overleftrightarrow{QR}$ and therefore has slope $\frac{b}{c}$. The altitude of the triangle from Q is perpendicular to $\overleftrightarrow{PR}$ and therefore has slope $\frac{a}{c}$. The equations of the altitudes from P and Q are respectively $bx - cy - ab = 0$ and $ax - cy - ab = 0$. Subtracting the second equation from the first, we obtain

$$\begin{array}{l} bx - cy - ab = 0 \\ \underline{ax - cy - ab = 0} \\ (b - a)x \qquad\quad = 0 \quad \Rightarrow x = 0 \text{ (since } a \neq b). \end{array}$$

Thus the point of intersection of the altitudes from P and from Q lies on the y-axis. Since the altitude on R is the y-axis, all the altitudes pass through the same point. Figure 12 illustrates the proof. ┘

Figure 12



The point of concurrency of the altitudes of a triangle is called the **orthocenter** of the triangle.

The above proof of Theorem 11.20 is analytic since it makes use of the number properties of the Cartesian plane. Let's pause for a minute to analyze it. In constructing the proof, we assumed the following definitions and properties of the Cartesian plane:

1. The slope of a line on two points (x_1, y_1) and (x_2, y_2) is given by $\frac{y_1 - y_2}{x_1 - x_2}$.
2. Perpendicular lines have slopes that are opposite reciprocals of one another (see Problem 4, Section 7.2.2).
3. The line with slope m that passes through the point (x_0, y_0) is given by the equation $y - y_0 = m(x - x_0)$, which can be written in standard form.
4. To determine the point of intersection of two lines, we can solve their equations simultaneously to obtain the ordered pair which represents the point of intersection.
5. The y-axis consists of the points in the plane with coordinates $(0, y)$ where y is a real number.

Using these definitions and properties, we then translated the geometrical data of the problem into algebraic terms so we could develop the proof in our model. That is, we constructed a triangle using the Cartesian coordinate system. We used the laws of algebra to transform the data and draw algebraic conclusions from them. That is, we wrote the equations of the lines and solved these equations simultaneously. Then we translated the results back to geometric language to obtain the desired result for Euclidean geometry.

Since Theorem 11.20 holds for a model of Euclidean geometry, it is also provable directly from the axioms we took in Sections 11.1.1 to 11.1.5. Such a proof is *synthetic*. For a synthetic proof of Theorem 11.20, we first state some familiar definitions, and lemmas (which we ask you to prove in the problems) that we need to use.

Definitions A **polygon** is a plane figure consisting of the points on n line segments (its **sides**), such that each side intersects exactly one other side at each of its endpoints (its **vertices**). A **quadrilateral** is a polygon with four sides. A **parallelogram** is a quadrilateral with both pairs of opposite sides parallel.

Lemma 1: The opposite sides of a parallelogram are congruent.

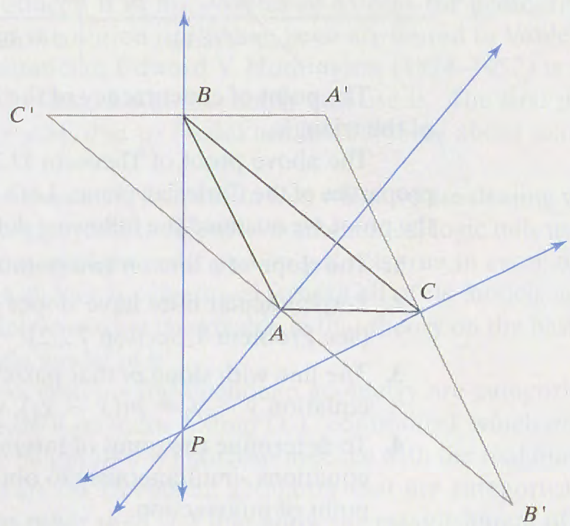
Lemma 2: The perpendicular bisectors of the sides of a triangle meet in a point. (This point is called the **circumcenter** of the triangle.)

With these we are now ready to prove Theorem 11.20 synthetically.

Theorem 11.20 The altitudes of a triangle are concurrent.

Proof (synthetic version): Given triangle ABC. Through A, draw the line parallel to $\overleftrightarrow{BC}$, through B draw the line parallel to $\overleftrightarrow{AC}$, and through C, draw the line parallel to $\overleftrightarrow{AB}$. We know these lines exist by the Euclidean parallel postulate. In this way we construct a new triangle, which we call $A'B'C'$ (Figure 13).

Figure 13



$ABA'C$ is a parallelogram, so be Lemma 1, $\overline{A'B} \cong \overline{AC'}$. Also, $BC'AC$ is a parallelogram, so $\overline{BC'} \cong \overline{AC}$. Thus $\overline{A'B} \cong \overline{BC'}$ (Axiom C-2). Since a line that is perpendicular to one of two parallel lines is perpendicular to the other (Problem 1 of Section 11.1.5), the altitude in ΔABC from B to $\overleftrightarrow{AC}$ is the perpendicular bisector of the side $\overline{A'C'}$ of $\Delta A'B'C'$. In similar fashion, the other two altitudes of ΔABC are the perpendicular bisectors of the other two sides of $\Delta A'B'C'$. By Lemma 2, the perpendicular bisectors of $\Delta A'B'C'$ are concurrent. Thus the altitudes of ΔABC are concurrent. ∎

You should compare the two proofs of Theorem 11.20—the analytic one constructed in the Cartesian model of Euclidean plane geometry, and the synthetic one deduced from the axioms and theorems of Euclidean plane geometry. The analytic proof requires a convenient position on the coordinate axis for the triangle. After that, the proof proceeds in a routine fashion, using the laws of algebra. The synthetic proof requires a little more ingenuity in the construction of the new triangle $A'B'C'$. Which method you prefer to employ is simply a matter of taste.

Another consequence from Euclidean geometry being a categorical system is that it enables us to determine whether a certain structure is a model of the Euclidean plane. One such structure is the Gaussian plane, named for Gauss, who studied it. The points in the **Gaussian plane** are the complex numbers $z = x + iy$, where x and y are real numbers, and i is $\sqrt{-1}$. A **Gaussian line** is the set of all points z that satisfy an equation $Dz + \overline{Dz} + E = 0$, where D is a nonzero complex number, $\overline{D}$ is its conjugate, and E is a real number.

We set up a 1-1 correspondence between the Cartesian plane and the Gaussian plane by mapping the real point (x, y) of the Cartesian plane into the complex point $x + iy$, and mapping the Cartesian line given by the equation $Ax + By + C = 0$ into the Gaussian line $Dz + \overline{Dz} + E = 0$, where $D = A - iB$, $\overline{D} = A + iB$, and $E = 2C$. We define the distance between two Gaussian points $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ to be $|z_1 - z_2| = |(x_1 - x_2) + i(y_1 - y_2)| = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$. With this definition, distance is preserved by the mapping. With these correspondences, the Cartesian plane and the Gaussian plane can be proved to be isomorphic, and as a result, the Gaussian plane is a model of Euclidean plane geometry. Thus, if a theorem can be proved in the Gaussian plane, then it is a theorem of Euclidean plane geometry.

We can also use a model to verify that a specific axiom (or its equivalent) is necessary to the definition of the theory in the presence of the other axioms, by showing that the reduced axiom set with that axiom removed does not define the categorical theory. For example, we can use the Cartesian model to investigate the role of Dedekind's axiom (D-1) in our axiom set. Dedekind's axiom enables us to obtain sufficient points on a line so the points in the Cartesian plane can be put into one-to-one correspondence with ordered pairs of real numbers. Suppose that we remove Dedekind's axiom from our set. One model of the resulting reduced axiom set is known as the *surd plane*. We define a *surd* as a real number x with the following property: We can calculate x by a finite number of additions, subtractions, multiplications, divisions, and extractions of square roots, starting with 0 and 1. Certainly, every rational number is a surd. However, not all real numbers are surds. Therefore, the lines of the surd plane are full of holes. Furthermore, the surd plane is countably infinite, while the Cartesian plane is uncountable, so there is no way to establish a one-to-one correspondence between the points and lines of the Cartesian plane and the points and lines of the surd plane.

We can therefore conclude that the axiom set of Sections 11.1.1 to 11.1.5 with D-1 removed produces two different (nonisomorphic) models, the Cartesian plane and the surd plane. A theory for which there are nonisomorphic models is **noncategorical**. Thus, our axiom set with Dedekind's axiom removed is noncategorical and cannot define the Euclidean geometry of our axiom set, which we deemed categorical.

Are axiom sets for the most familiar mathematical structures categorical? No. For instance, the rational numbers and the real numbers with the usual operations of addition and multiplication satisfy the axioms for a field, but we cannot set up a one-to-one correspondence between these two models. Thus a theorem true for the rational numbers may not be true for the real numbers, and vice versa. But the axioms of a complete ordered field are categorical. So every theorem that holds for a complete ordered field also holds for the real numbers. (See Section 2.3.1.)

We have now come to the end of this book, but in a way we have returned to its beginning. Recall that in Chapter 2 we used the number line as a *geometric* model of the real numbers and as inspiration for properties of real numbers. In this unit, we have shown how real numbers enable us to develop an *algebraic* model for geometry. These models demonstrate the cross-fertilizations of algebraic and geometric ideas that permeate all of mathematics and lead to the richness of its theory and applications.

11.2.2 Problems

1. Prove Lemma 1: The opposite sides of a parallelogram are congruent. You may choose a synthetic or analytic proof.

2. Prove Lemma 2 both synthetically and analytically: The perpendicular bisectors of the sides of a triangle meet in a point, which is called the circumcenter of the triangle.

3. Prove Theorem 11.20 from Lemma 2, by means of the transformations discussed in Chapter 8. (*Hint*: Construct a size change that maps the altitudes of any triangle onto the perpendicular bisectors of the sides of a second triangle.)

4. Prove the following theorem synthetically and analytically: The diagonals of a parallelogram bisect each other.

5. Prove analytically and synthetically that any three non-collinear points lie on a circle.

6. An axiom is called **independent** in a set S of axioms if it cannot be proved or disproved using only the other axioms in S. An axiom system S is **complete** if it is impossible to add an independent axiom to it. By an indirect argument, prove that if an axiom system S is categorical, then it is complete.

7. In Section 11.1.1 we discussed a 7-point geometry that satisfies the incidence axioms. Our discussion was based on an algebraic model of the geometry. The following axioms *define* the geometry of that model (Fano's geometry):

F-1: There exists at least one line.

F-2: Every line has exactly three points on it.

F-3: Not all points are on the same line.

F-4: For each two distinct points, there exists exactly one line on both of them.

F-5: For each two lines there exists at least one point on both of them.

a. Using the axioms, prove that there are exactly 7 points in Fano's geometry, and therefore in every model of it. *Note*: You must not only show 7 points exist. You must also show an 8th point cannot exist.

b. Create a model for Fano's geometry that is different from the one given in Section 11.1.1.

c. Show that axiom F-5 of Fano's geometry is independent of the other four axioms. (*Hint*: Create a model for a geometry that satisfies all the axioms for Fano's geometry except for axiom F-5.)

d. Do you think Fano's geometry as defined by the above axioms is categorical? Explain why or why not.

Chapter Projects

1. Young's geometry.

a. Construct an axiom set (Y-1 to Y-5) for Young's geometry of Problem 9 in Section 11.1.1. Then prove that the incidence axioms (I-1 to I-4) are satisfied directly from the axioms you create.

b. Consider the following proof that every point in Young's geometry is on at least four lines.

Proof: Let P be any point, and let ℓ be any line that does not contain P. ℓ contains exactly three points. P and each point on ℓ must determine a distinct line; therefore, we have at least three lines. Now, there must be a line that contains P but contains no points on ℓ. Therefore, we have at least four lines on P. The theorem is proved.

Refer to the axioms you created in part **a** and justify every step of the above proof by means of an axiom. Then use your axioms to prove that Young's geometry must contain exactly nine points and exactly twelve lines.

c. Alter your axioms of part **a** for Young's geometry to assert that instead of three points, there are exactly two points on every line. How many points and lines would the new geometry have?

d. Alter your axioms for Young's geometry to assert that there are exactly four points on every line. How many points and lines would this new geometry have?

e. Generalize your results from parts **c** and **d** for the case where each line contains exactly n points ($n =$ some positive integer). That is, consider a finite model satisfying the axioms for Young's geometry, assuming that instead of three points on a line, there are exactly n points on every line. How many points and lines would the geometry have?

f. Define parallel lines as lines with no points in common. Prove that two lines parallel to a third line are parallel to each other. Use the axioms to construct your proof. (*Hint*: First verify the statement by observing the model. Then construct your proof by assuming that two lines parallel to a third line are not parallel to each other and show that this assumption leads to a contradiction.)

2. Pappus's theorem. The following theorem was discovered and proved by Pappus of Alexandria about A.D. 340. If A, B, and C are three distinct points on one line, and if A', B', and C' are three different distinct points on a second line, then the intersections of $\overline{AC'}$ and $\overline{CA'}$, $\overline{AB'}$ and $\overline{BA'}$, and $\overline{BC'}$ and $\overline{CB'}$ are collinear.

a. Prove that Pappus's theorem holds in Young's geometry (Problem 9 of Section 11.1.1) for the six points on any pair of parallel lines.

b. It is possible to create a finite geometry from a geometric figure. For example, the finite geometry of Pappus arises from a figure resulting from his theorem. The figure, called the **Pappus configuration**, consists of nine points and nine lines, and inspires the definition of a new finite geometry with the following axioms:

P-1. There exists at least one line.

P-2. Every line has exactly three points.

P-3. Not all points are on the same line.

P-4. If a point is not on a given line, then there exists exactly one line on the point that is parallel to the given line.

P-5. If P is a point not on a line, then there exists exactly one point P' on the line such that no line joins P and P'.

P-6. With the exception of Axiom P-5, if P and Q are distinct points, then exactly one line contains both of them.

Prove the following theorems in this geometry:

a. Each point in the geometry of Pappus lies on exactly three lines.

b. There are exactly three lines on each point.

c. The geometry of Pappus has nine points and nine lines.

3. George E. Martin, in Chapter 5 of *The Foundations of Geometry and the Non-Euclidean Plane* (Springer-Verlag, New York, 1975), gives several models of Euclidean plane geometry. Choose one of these models, and write an essay describing it, showing how it is essentially the same as the Cartesian plane.

4. Investigate the work of Fano, in particular his three-dimensional finite incidence geometry. Design a talk for a local high school mathematics club, telling them about this mathematician and describing some characteristics of his three dimensional model.

5. Explore the interesting connection between Fano planes and Hamming error-correcting codes in *A Course in Modern Geometries* by J. Cederberg (New York: Springer-Verlag, 1989) or in *The Mathematical Theory of Coding* by I. F. Blake and R. C. Mullin (New York: Academic Press, 1975) or in texts in modern applied algebra. Prepare a lesson demonstrating this application of finite geometry.

6. The first model of plane hyperbolic geometry was given by the Italian mathematician Eugenio Beltrami in 1868. It is called a *pseudosphere* and is a surface formed by revolving a curve, called the *tractrix*, about an asymptote. Search the Internet or consult some books on non-Euclidean geometry for information about the pseudosphere. Draw a geometric model of the pseudosphere, and determine what are the points and lines of this model.

7. M. C. Escher's circular woodcuts (Circle Limits I, II, III, and IV) depict similar shapes (fish, angels, etc.) that diminish in size as they recede from the center and fit together to fill and cover a disk. I, II, and IV are based on a model of the hyperbolic plane that is owed to Henri Poincaré. Do some research on this and write an essay discussing how the Poincaré model was used in Escher's Circle Limit prints. Consult, for example, H. S. M. Coxeter; "The Trigonometry of Escher's Woodcut 'Circle Limit III'," *Mathematical Intelligencer* 18(4), 1996, or M. C. Escher: *Art and Science*, North-Holland, 1986 (edited by Coxeter and others).

8. Read the article "From Pappus to Today: The History of a Proof," *Mathematical Gazette* 74, 1990, 6–11, by Michael Deakin, which explores different proofs of the theorem that base angles of an isosceles triangle are equal. Select the proof that you believe has the greatest pedagogical advantage, and describe why you believe this to be so.

Bibliography

Unit 11.1 References

Prenowitz, Walter, and Meyer Jordan. *Basic Concepts of Geometry*. New York: Blaisdell, 1965.
The authors explore Euclidean and non-Euclidean geometries, as well as finite geometries, using incidence as a basic unifying idea.

Rosenfeld, B. A. *History of Non-Euclidean Geometry*. New York: Springer-Verlag, 1988.
This book investigates the mathematical and philosophical factors underlying the discovery of non-Euclidean geometry and the extension of the concept of space.

Unit 11.2 References

Henderson, David. *Experiencing Geometry on Plane and Sphere*. Upper Saddle River, NJ: Prentice Hall, 1996.
The author gives a series of problems designed to promote the learning of geometry by means of intuition and reasoning from experience.

Martin, George E. *The Foundations of Geometry and the Non-Euclidean Plane*. New York: Springer-Verlag, 1975.
The author provides an axiomatic development of the Euclidean and hyperbolic planes, providing historical aspects. Axiomatic systems of Euclid, Hilbert, and Pieri are discussed.

Millman, R. S., and G. D. Parker. *Geometry: A Metric Approach with Models*. New York: Springer-Verlag, 1981.
This book seeks to give the reader some intuition about Euclidean and hyperbolic geometry through an examination of analytic models.

Moise, Edwin E. *Elementary Geometry from an Advanced Standpoint*. Reading, MA: Addison-Wesley, 1963.
This classic geometry textbook gives a rigorous development of Euclidean and hyperbolic geometries, and attempts to reexamine familiar topics, providing valid definitions and proofs for concepts and theorems treated in more elementary courses.

Smith, James T. *Methods of Geometry*. New York: John Wiley & Sons, Inc., 2000.
This book builds on the mathematical knowledge of high school and lower division university courses to explore two- and three-dimensional geometry. It includes a wealth of nonroutine exercises and projects, an extensive annotated bibliography, and historical references.

Unit 11.3 References

Greenberg, Marvin J. *Euclidean and Non-Euclidean Geometries*. Third Edition. New York: W. H. Freeman and Co., 1993.
The author provides a rigorous treatment of the foundations of Euclidean geometry and an introduction to hyperbolic geometry, emphasizing its Euclidean models. This edition includes a discussion of the history of Euclid's parallel postulate, and the birth of non-Euclidean geometry.

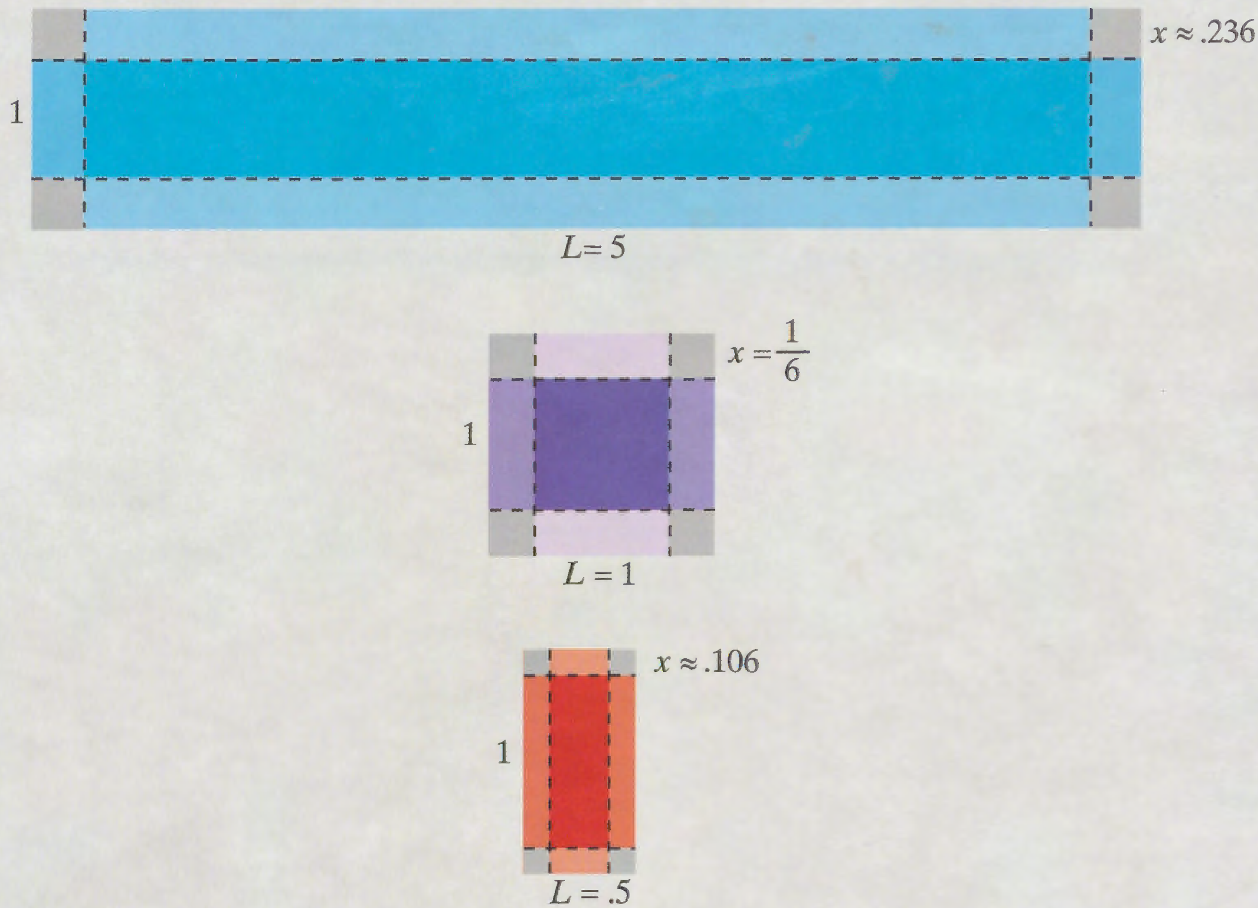
Hartshorne, Robin. *Geometry: Euclid and Beyond*. New York: Springer-Verlag, 2000.
This book examines Euclid's *Elements*, explores questions that arise from it, and gives modern answers.

INDEX

(page numbers in **boldface** contain a definition of the term)

O

P



Pearson Education

Prentice Hall
Upper Saddle River, NJ 07458
www.prenhall.com/statistics

ISBN 0-13-044941-5
90000
9 780130 449412