Pre-Algebra [DRAFT]

H. Wu

April 21, 2010; *revised,* October 26, 2011

General Introduction

The content of these notes is the mathematics that is generally taught in grades 6–8. This is a no frills, bare essentials course for helping you to teach mathematics in the middle school classroom, and is *not* designed to show you how mathematics, deep down, is just lots of fun. We review most of the standard topics of the middle school mathematics curriculum. *However*, the presentation of this material in the standard textbooks, be they traditional or reform, is riddled with mathematical errors. What is presented in this institute, while bearing superficial resemblances to what you normally find in textbooks, will likely be very different in terms of precision, sequencing, and reasoning. You will probably have to rethink some of this material even if you believe you already know them very well.

Let us look at the concept of congruence, a main point of emphasis in these notes. Most textbooks would have you believe that it means **same size and same shape**. As mathematics, this is totally unacceptable, because "same size" and "same shape" are words that mean different things to different people, whereas mathematics only deals with clear and unambiguous information. I will therefore suggest that you approach the teaching of this concept completely differently. First make sure that *you* know what reflections, translations and rotations are, then devise hands-on activities for your students to familiarize them with these concepts, and finally, teach them that two sets are congruent if one can carry one set onto the other by use of a finite number of reflections, translations and rotations.

You see right away that we will be doing standard middle school mathematics, but for a change, we will do it in a way that is consistent with how mathematics is supposed to be done. The hope is that by the time you are finished with these notes, you will begin to recognize school mathematics as a coherent subject with every concept and skill placed in a logically correct hierarchy. If I may express this idea by use of an analogy, it would be like bringing bookshelves to a roomful of books scattered all over the floor and trying to put the books on the shelves using a well-understood organizing principle. Once arranged this way, any book in the room can be easily accessed in the future. Likewise, if we can re-organize mathematical thoughts logically in our mind, we can much more easily access and make use of them.

But why? An obvious reason is that if we want students to see mathematics as a tool to help solve problems, the different parts of this tool must be freely accessible. This cannot happen if we as teachers do not have free access to these parts ourselves. A less obvious, but perhaps more compelling reason is that teaching mathematics merely as a jumbled collection of tasks has led our nation to a severe mathematics education crisis.[1] It would be reasonable to attribute a good deal of students' non-learning of mathematics to their being fed such jumbled information all the way from kindergarten to grade 12.[2] These notes are dedicated to making improvements in mathematics instructions, one classroom at a time.

The main goal of these notes is to provide the necessary background for the teaching of algebra. Getting all students to take algebra around grade 8 is at present a national goal. For an in-depth discussion, see the National Mathematics Panel's Conceptual Knowledge and Skills Task Group Report:

http://www.ed.gov/about/bdscomm/list/mathpanel/report/conceptual-knowledge.pdf

Currently, most school students are deficient in their knowledge of the two pillars that support algebra: rational numbers and similar triangles; these two topics are the subject of five of the seven chapters in these notes. In the current school curriculum, one does not associate the learning of similar triangles as a pre-requisite to the learning of algebra. *But it is*, and this failure to give adequate support to our students' learning of algebra is one of the flaws in mathematics instructions that we set out to remedy.

Overall, these notes will strive to improve mathematics teaching by emphasizing, throughout, the following three principles:

(I) *Precise definitions are essential.* Definitions are looked upon with something close to disdain by most teachers because "they are nothing more than something to be memorized". Such an attitude stems from poor professional development that breeds such a misconception of mathematics. First of all, memorizing important facts

[1] See, for example, *Rising Above the Gathering Storm*, The National Academies Press, 2007. Also, http://www.nap.edu/catalog.php?record_id=11463. Or *Foundations for Success: The National Mathematics Advisory Panel Final Report*, U.S. Department of Education, 2008. Also, http://www.ed.gov/about/bdscomm/list/mathpanel/reports.html.

[2] Such a statement should not be misinterpreted to mean that this is the only reason for students' non-learning. There is enough blame to go around.

is an integral part of life (you memorize your name, your home address, your cell phone, the password of your computer account, etc.), and you will have to memorize all the definitions we use. No apology will be offered. But the idea that a definition in mathematics is nothing but "one more thing to memorize" must be eradicated. In mathematics, precise definitions are the bedrock on which all logical reasoning rests because mathematics does not deal with vaguely conceived notions. These notes will respect this fundamental characteristic of mathematics by offering precise definitions for many concepts in the school curriculum usually used with no definitions: *fraction, decimals, sum of fractions, product of fractions, ratio, percent, polygon, congruence, similarity, length, area,* etc.

(II) *Every statement should be supported by reasoning.* There are no unexplained assertions in these notes. If something is true, a reason will be given. Although it takes some effort to learn the logical language used in mathematical reasoning, in the long run, the presence of reasoning in all we do eases the strain of learning and disarms disbelief. It also has the salutary effect of putting the learner and the teacher on the same footing, because the ultimate arbiter of truth will no longer be the teacher's authority but the compelling rigor of the reasoning.

(III) *Mathematics is coherent.* You will see that these notes unfold *logically* and *naturally* rather than by fits and starts. On the one hand, each statement follows logically from the preceding one, and on the other, the various statements form parts of an unending story rather than a disjointed collection of disparate tricks and factoids. A striking example of the failure of coherence is the common explanation of the theorem on equivalent fractions, which states that $\frac{m}{n} = \frac{km}{kn}$ *for all fractions* $\frac{m}{n}$ *and for all positive integer* k. Most book would have you believe that this is true because

$$\frac{m}{n} \;=\; 1 \times \frac{m}{n} \;=\; \frac{k}{k} \times \frac{m}{n} \;=\; \frac{km}{kn}$$

Unfortunately, the last step depends on knowing how to multiply fractions. But the definition of fraction multiplication is the most subtle among the four arithmetic operations on fractions, whereas the theorem on equivalent fractions should be proved as soon as fractions are defined. To invoke something not yet explained and technically more complex to explain something logically simpler and more elementary is a blatant violation of the fundamental structure of mathematics. Such violations abound in current textbooks.

An example of the interconnections among seemingly different topics that hold the subject together is the fact that the concept of similarity in Chapter 6 relies on a knowledge of dividing fractions (Chapter 1) and congruence (Chapter 5). Another example is the key role played by congruence in the considerations of length, area, and volume (Chapter 7). And as a final example, you will notice that the division of whole numbers, of fractions (Chapter 1), and of rational numbers (Chapter 2) are conceptually identical.

I hope you find that these notes make more sense of the mathematics you know because they observe these basic principles. As far as this institute is concerned, however, what matters is that you can translate this new-found knowledge into better teaching in your classroom. I am counting on you to make this next step.

I am very grateful to Larry Francis for correcting a large number of linguistic infelicities and misprints.

Suggestions on How to Read These Notes

The major conclusions in these notes, as in all mathematics books, are summarized into **theorems**; depending on the author's (and other mathematicians') whims, theorems are sometimes called **propositions**, **lemmas**, or **corollaries** as a way of indicating which theorems are deemed more important than others (note that a formula or an algorithm is just a theorem). This idiosyncratic classification of theorems started with Euclid around 300 B.C. and it is too late to change now. The main concepts of mathematics are codified into **definitions**. Definitions are set in **boldface** in these notes when they appear for the first time. A few truly basic ones among the definitions are even individually displayed in a separate paragraph, but most of the definitions are embedded in the text itself. Be sure to watch out for them.

The statements of the theorems as well as their proofs depend on the definitions, and proofs (= reasoning) are the guts of mathematics.

A preliminary suggestion to help you master the content of these notes is for you to

> copy out the statements of every definition, theorem, proposition, lemma, and corollary, along with page references so that they can be examined in detail if necessary,

and also to

> summarize the main idea of each proof.

These are good study habits. When it is your turn to teach your students, be sure to pass on these suggestions to them. A further suggestion is that you might consider posting some of these theorems and definitions in your classroom.

You should also be aware that reading mathematics is not the same as reading a gossip magazine. You can probably flip through such a magazine in an hour, if not less. But in these notes, there will be many passages that require careful reading and re-reading, perhaps many times. I cannot single out those passages for you because they will be different for different people. We don't all learn the same way. What is true under all circumstances is that you should accept as a given that mathematics books make for exceedingly slow reading. I learned this very early in my career. On my very first day as a graduate student many years ago, a professor, who was eventually to become my thesis advisor, was lecturing on a particular theorem in a newly published volume. He mentioned casually that in the proof he was going to present, there were two lines in that book that took him fourteen hours to understand and he was going to tell us what he found out in those long hours. That comment greatly emboldened me not to be afraid to spend a lot of time on any passage in my own reading.

If you ever get stuck in any passage of these notes, take heart, because that is nothing but par for the course.

Chapter 1: Fractions

In this and the next chapter, we are going to develop a theory of fractions and rational numbers (positive and negative fractions) that is suitable for use in upper elementary and middle school. Every teacher must have a firm grasp of fractions and rational numbers, because **school mathematics as a whole is about rational numbers.** These two chapters give an exposition of rational numbers that is suitable for use in a classroom of grades 5–7.

In reading these two chapters, please keep in mind that the emphasis here will not be on individual facts or skills. This is not to say that facts and skills are not important, — they are — but, assuming that you are somewhat familiar with them, we will be more concerned with the *reorganization* of these facts and skills so that they form a logical and coherent whole that is compatible with the learning processes of upper elementary and middle school students. The hope is that, with this reorganization, you as a teacher will be able to explain fractions to your students in a way that makes sense to them and to you yourself. This is the first step toward establishing mathematical communication between you and your students. For example, it may come as a surprise to you that it is possible to develop the concept of adding fractions without once mentioning "the least common denominator", and that the invert-and-multiply rule for the division of fractions is a *theorem* that can be proved on the basis of a precise definition of fraction division.

The teaching of fractions is the most problematic part of school mathematics because, in the usual way it is done, there are hardly any valid definitions offered and almost nothing is ever explained. The resulting non-learning of fractions is not only a national scandal within the state of mathematics education, but also a major stumbling block in students' learning of algebra. The critical importance of fractions to the learning of algebra is beginning to be recognized. See, for example, the report of the National Mathematics Panel (2008),

http://www.ed.gov/about/bdscomm/list/mathpanel/report/conceptual-knowledge.pdf

See also the article, H. Wu, *From arithmetic to algebra*,

http://math.berkeley.edu/~wu/C57Eugene_2.pdf

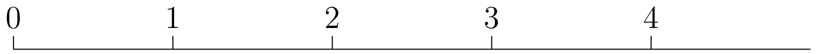
For example, can you recall from your own K-12 experience if you were ever told what it means to multiply $\frac{5}{4} \times \frac{3}{7}$ and, moreover, why that is equal to $\frac{5\times 3}{4\times 7}$? If not, then it should give you incentive to do better when it is your turn to teach. What you will learn in these two chapters, and perhaps all through these notes, will be a reorganization of the bits and pieces that you learned haphazardly in K–12 into a coherent body of knowledge. Your job will be to make this re-organized knowledge accessible to your students. You are being asked to become an advocate of teaching school mathematics the way mathematics *should* be taught: giving precise definitions to all the concepts and explaining every algorithm and every skill. The purpose of these notes is to optimize your chances of success in this undertaking.

1 Definition of a fraction

Mise-en-scène

Mathematics rests on precise definitions. We need a definition of a fraction, not only because this is what mathematics demands, but also because children need a precise mental image for fractions to replace the mental image of their fingers for whole numbers.

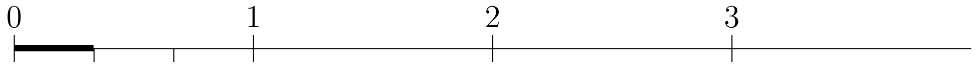
We begin with the *number line*. On a line which is (usually chosen to be) horizontal, we pick a point and designate it as 0. We then choose another point to the right of 0 and, by reproducing the distance between 0 and this point, we get an infinite sequence of equi-spaced points to the right of 0. Think of this as an infinite ruler. Next we label all these points by the nonzero whole numbers 1, 2, 3, . . . in the usual manner. Thus all the whole numbers $\mathbf{N} = \{0, 1, 2, 3, \ldots\}$ are now displayed on the line as equi-spaced points increasing to the right of 0, as shown:



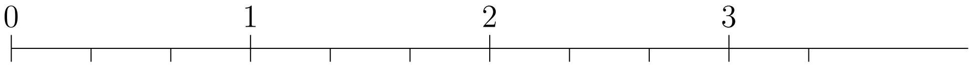
A horizontal line with an infinite sequence of equi-spaced points identified with **N** on its right side is called the **number line**. By definition, a **number** is just a point on the number line. Note that except for the original sequence of equi-spaced points which we have chosen to denote by 1, 2, 3, etc., most numbers do not have recognizable names as yet. The next order of business will be to name more numbers, namely, the *fractions*.

Now fractions have already been introduced to students in the primary grades, and their basic understanding of fractions is that they are "parts of a whole". The transition from "parts of a whole" to "a point on the number line" has to be handled with care. This is because this transition, which should occur in the fifth or sixth grade, is students' first serious introduction to abstract thinking in mathematics, and it is anything but natural to them. After all, "parts of a whole" is an *object*, e.g., an area, a part of a pizza, an amount of water in a glass, or a certain line segment, but not a point on a line. Therefore, the following ***informal*** discussion is intended to smooth out this transition as well as prepare you for the contingency of having to convince your students to accept a fraction as a certain point on the number line.

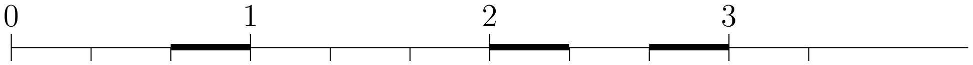
We begin the informal discussion by considering a special case: how the fractions with denominator equal to 3, i.e., $\frac{1}{3}$, $\frac{2}{3}$, $\frac{3}{3}$, $\frac{4}{3}$, *etc., come to be thought of as a certain collection of points on the number line. We take as our "whole" the unit segment* $[0, 1]$. (We will denote the segment from c to d, with $c < d$, by $[\boldsymbol{c}, \boldsymbol{d}]$.) *The fraction* $\frac{1}{3}$ *is therefore one-third of the whole, i.e., if we divide* $[0, 1]$ *into 3 equal parts,* $\frac{1}{3}$ *stands for one of the parts. One obvious example is the thickened segment below:*



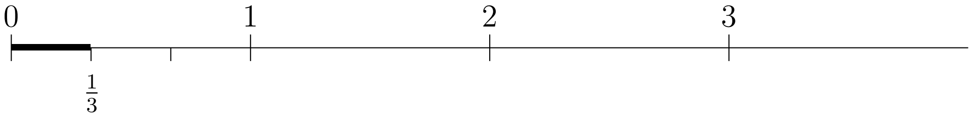
Of course this particular thickened segment is not the only example of "a part when the whole is divided into 3 equal parts". Let us divide, not just $[0, 1]$, *but every segment between two consecutive whole numbers —* $[0, 1]$, $[1, 2]$, $[2, 3]$, $[3, 4]$, *etc. — into three equal parts. Then these division points, together with the whole numbers, form an infinite sequence of equi-spaced points, to be called* **the sequence of thirds***:*



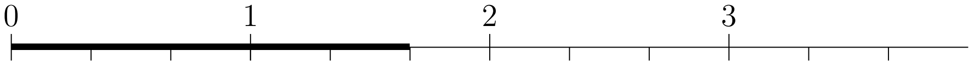
For convenience, we call any segment between consecutive points in the sequence of thirds a **short segment***. Then any of the following thickened short segments is "one part when the whole is divided into 3 equal parts" and is therefore a legitimate representation of* $\frac{1}{3}$*:*

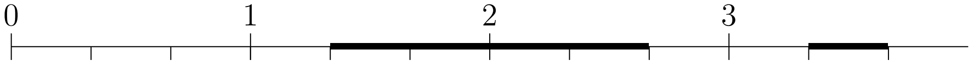


The existence of these multiple representations of $\frac{1}{3}$ *complicates life and prompts the introduction of the following* **standard representation** *of* $\frac{1}{3}$*, namely, the short segment whose left endpoint is 0 (see the very first example above of a thickened short segment). With respect to the standard representation of* $\frac{1}{3}$*, we observe that this short segment determines its right endpoint, and vice versa: knowing this segment means knowing its right endpoint, and knowing the right endpoint means knowing this segment. In other words, we may as well identify the standard representation of* $\frac{1}{3}$ *with its right endpoint. It is then natural to denote this right endpoint by* $\frac{1}{3}$*:*

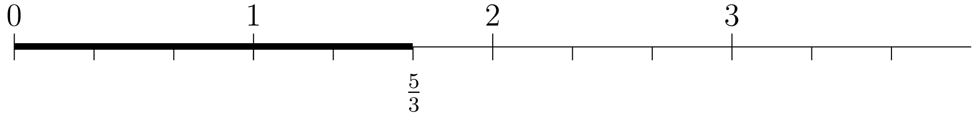


In like manner, by referring to the sequence of thirds and its associated short segments, the fraction $\frac{5}{3}$*, being 5 of these short segments, can be represented by any of the following collections of thickened short segments:*

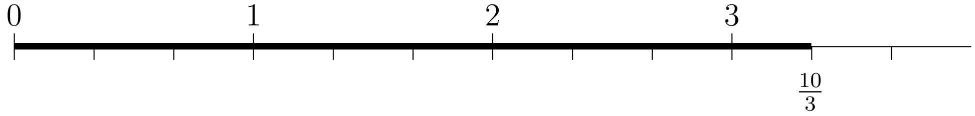




Again, our **standard representation** *of $\frac{5}{3}$ is the first one, which consists of 5 adjoining short segments abutting 0. This standard representation is completely determined by its right endpoint, and vice versa. Thus to specify the standard representation of $\frac{5}{3}$ is to specify its right endpoint. For this reason, we identify the standard representation of $\frac{5}{3}$ with its right endpoint, and proceed to denote the latter by $\frac{5}{3}$, as shown.*

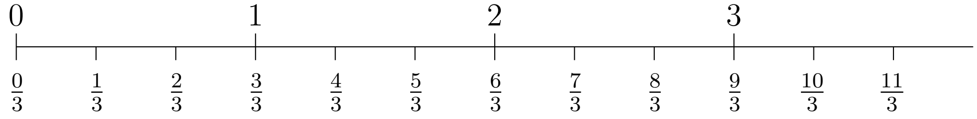


In general then, a fraction $\frac{m}{3}$ (where m is some whole number) has the **standard representation** *consisting of m adjoining short segments abutting 0, where* “short segment” *refers to a segment between consecutive points in the sequence of thirds. Since we may identify this standard representation of $\frac{m}{3}$ with its right endpoint, we denote the latter simply by $\frac{m}{3}$. The case of $m = 10$ is shown below:*



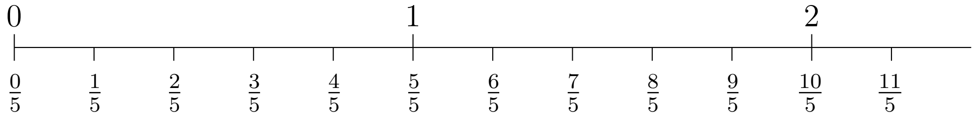
We note that in case $m = 0$, $\frac{0}{3}$ is just 0.

Having identified each standard representation of $\frac{m}{3}$ with its right endpoint, each point in the sequence of thirds now acquires a name, as shown below. These are exactly the fractions with denominator equal to 3.



In terms of the sequence of thirds, each fraction $\frac{m}{3}$ is easily located: the point $\frac{m}{3}$ is the m-th point to the right of 0. *Thus if we ignore the denominator, which is 3, then* the naming of the points in the sequence of thirds is no different from the naming of the whole numbers.

Of course the consideration of fractions with denominator equal to 3 extends to fractions with other denominators. For example, replacing 3 by 5, then we get **the sequence of fifths***, which is a sequence of equi-spaced points obtained by dividing each of* $[0,1]$, $[1,2]$, $[2,3]$, $\ldots$, *into 5 equal parts. The first 11 fractions with denominator equal to 5 are now displayed as shown:*



Finally, if we consider all the fractions with denominator equal to n*, then we would be led to* **the sequence of n-ths***, which is the sequence of equi-spaced points resulting from dividing each of* $[0,1]$, $[1,2]$, $[2,3]$, $\ldots$, *into* n *equal parts. The fraction* $\frac{m}{n}$ *is then the* m*-th point to the right of 0 in this sequence.*

This ends the informal discussion.

The formal definition

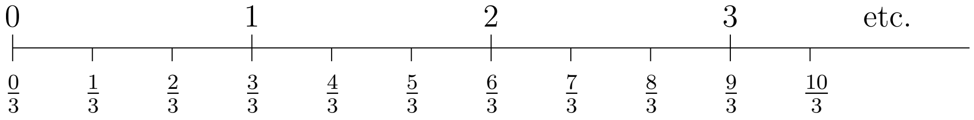
We now turn to the formal definition of a fraction.

We will begin by making precise the common notion of "equal parts". A segment $[a,b]$ is said to be of **length k** for a number k if, when we slide $[a,b]$ along the number line until a is at 0, the right endpoint b lies over the number k.[3] In particular, the unit segment has length 1. We say a segment $[a,b]$ **is divided into m equal parts** if $[a,b]$ is expressed as the union of m adjoining, nonoverlapping segments of equal length. A sequence of points is said to be **equi-spaced** if the segments between consecutive points in the sequence are all of the same length.

Divide each of the line segments $[0,1]$, $[1,2]$, $[2,3]$, $[3,4]$, $\ldots$, into 3 equal parts. The totality of division points, which include the whole numbers, form a sequence

[3] It will be observed that, insofar as the only numbers (i.e., points on the number line) which have been given names thus far are the whole numbers, this length k may not have any names we recognize. But as we proceed to name the fractions, there will be more lengths that we will recognize as fractions.

of equi-spaced points, to be called **the sequence of thirds**. By definition, **the fraction $\frac{1}{3}$** is the first point in the sequence to the right of 0, $\frac{2}{3}$ is the second point, $\frac{3}{3}$ is the third point, and in general, $\frac{m}{3}$ is the m-th point in the sequence to the right of 0, for any nonzero whole number m. By convention, we also write 0 for $\frac{0}{3}$. Note that $\frac{3}{3}$ coincides with 1, $\frac{6}{3}$ coincides with 2, $\frac{9}{3}$ coincides with 3, and in general, $\frac{3m}{3}$ coincides with m for any whole number m. Here is the picture:



The fraction $\frac{m}{3}$ is called the ***m*-th multiple** of $\frac{1}{3}$. Note that the way we have just introduced the multiples of $\frac{1}{3}$ on the number line is exactly the same way that the multiples of 1 (i.e., the whole numbers) were introduced on the number line. In other words, if we do to $\frac{1}{3}$ exactly what we did to the number 1 in putting the whole numbers on the number line, then we would also obtain every $\frac{m}{3}$ for a whole number m.

In general, if a nonzero $n \in \mathbf{N}$ is given, we introduce a new collection of points on the number line in the following way: Divide each of the line segments $[0, 1]$, $[1, 2]$, $[2, 3]$, $[3, 4]$, ... into n equal parts, then these division points (which include the whole numbers) form an infinite sequence of equi-spaced points on the number line, to be called **the sequence of *n*ths**. The first point in the sequence to the right of 0 is denoted by $\frac{1}{n}$, the second point by $\frac{2}{n}$, the third by $\frac{3}{n}$, etc., and the mth point in the sequence to the right of 0 is denoted by $\frac{m}{n}$. By convention, $\frac{0}{n}$ is 0.

Definition *The collection of all the sequences of $n$ths, as n runs through the nonzero whole numbers 1, 2, 3, ..., is called the* **fractions**. *The mth point to the right of 0 in the sequence of $n$ths is denoted by $\frac{m}{n}$. The number m is called the* **numerator** *and n, the* **denominator** *of the symbol $\frac{m}{n}$. By the traditional abuse of language, it is common to say that m and n are the* **numerator and denominator, respectively, of the fraction $\frac{m}{n}$**.[4] *By definition,* 0 *is denoted by* $\frac{0}{n}$ *for any n.*

[4] The correct statement is of course that "m is the numerator of the symbol which denotes the fraction that is the mth point of the sequence of $n$ths, and n is the denominator of this symbol." (Needless to say, it takes talent far above the norm to talk like this.)

Remark All the care that goes into this definition of a fraction is not an empty exercise in formalism, much less "another fact to memorize", which is a common misconception of what a definition is. What this definition does is to set in motion how the rest of this chapter will unfold, namely, if any assertion is made about fractions, that assertion must be explained (i.e., proved) by referring to this meaning of a fraction, no more and no less. This is how seriously you must take this and any other definition. So ***memorize the definition of a fraction any way you can***, because you must have instant recall of this definition at all times.

By tradition, a fraction $\frac{m}{n}$ so that $m < n$ is called **proper**, whereas it is **improper** if $m \geq n$. As before, we shall refer to $\frac{m}{n}$ as the **mth multiple of $\frac{1}{n}$.** In the future, *we will relieve the tedium of always saying the denominator n of a fraction $\frac{m}{n}$ is nonzero by simply* ***not*** *mentioning it.*

Some special features

A few remarks about the definition of a fraction are in order:

(A) In general, if m is a multiple of n, say $m = kn$,[5] then it is self-evident that $\frac{n}{n} = 1$, $\frac{2n}{n} = 2$, $\frac{3n}{n} = 3$, $\frac{4n}{n} = 4$, and in general,

$$\frac{kn}{n} = k, \quad \text{for all whole numbers } k,\ n, \text{ where } n > 0.$$

In particular,

$$\frac{m}{1} = m \qquad \text{and} \qquad \frac{m}{m} = 1$$

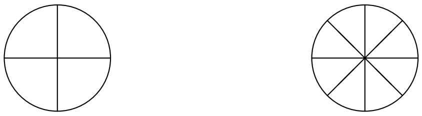
for any whole number m.

(B) For the study of fractions, the unit is of extreme importance. ***On the number line, it is impossible to say which point is what fraction until the number 1 has been fixed.*** This means that in the classroom, students need to be reminded that every fraction has to refer to a unit: does $\frac{1}{3}$ mean a third of the *volume* of the liquid in a conical container or a third of this liquid by *height*? Or, a fraction

[5] We follow the standard convention of suppressing the multiplication symbol × between two letter symbols.

$\frac{5}{7}$ could be five-sevenths of a bucket of water by volume or five-sevenths in dollars of your life-savings.

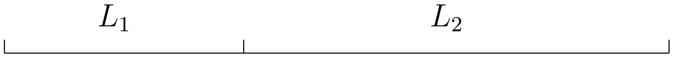
(C) A unit must be described with precision. An example of a common error is *to refer to a pizza as the unit* ("the whole"), and ask what fraction is represented by putting one of the four pieces on the left together with one of the eight pieces on the right:



Many students will not come up with the expected answer of $\frac{3}{8}$, because they take 1 to be the *shape* of the pizza and they don't know how to put *two shapes together* to get a fraction. Better to tell them that 1 represents *the area of the pizza.*

(D) We have been talking about *the* number line, but in a literal sense this is wrong. A different choice of the line or even a different choice of the positions of the number 0 and 1 would lead to a different number line. What is true, however, is that anything done on one number line can be done on any other *in exactly the same way.* In technical language, all number lines are *isomorphic*,[6] and therefore we identify all of them. Now it makes sense to speak of *the* number line.

(E) Although a fraction is formally a *point* on the number line, the informal discussion above makes it clear that on an intuitive level, a fraction $\frac{m}{n}$ is just the segment $[0, \frac{m}{n}]$. So in the back of our minds, the *segment* image never goes away completely, and this fact is reflected in the language we now introduce. First, a definition: the **concatenation** of two segments L_1 and L_2 on the number line is the line segment obtained by putting L_1 and L_2 along the number line so that the right endpoint of L_1 coincides with the left endpoint of L_2:



[6] Corresponding to the fact that all complete ordered fields are isomorphic.

Thus the segment $[0, \frac{m}{n}]$ is the concatenation of exactly m segments each of length $\frac{1}{n}$, to wit, $[0, \frac{1}{n}]$, $[\frac{1}{n}, \frac{2}{n}]$, $\ldots$, $[\frac{m-1}{n}, \frac{m}{n}]$. Because we identify $[0, \frac{m}{n}]$ with the point $\frac{m}{n}$, and $[0, \frac{1}{n}]$ with $\frac{1}{n}$, it is natural to adopt the following suggestive terminology to express the fact that *the segment* $[0, \frac{m}{n}]$ *is the concatenation of exactly* m *segments each of length* $\frac{1}{n}$:

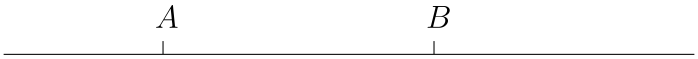
$$\boldsymbol{\frac{m}{n}} \ \textbf{is} \ \boldsymbol{m} \ \textbf{copies of} \ \boldsymbol{\frac{1}{n}}$$

(F) In education research, the meaning of the equal sign is a subject that is much discussed, mainly because the meaning of *equality* is never made clear in school mathematics. For this reason, we make explicit the fact that two fractions (which are two points on the number line) are said to be **equal** if they are the same point. If the given fractions are denoted by $\frac{k}{\ell}$ and $\frac{m}{n}$, then we denote the equality by the usual symbol:

$$\boldsymbol{\frac{k}{\ell} = \frac{m}{n}}$$

We have already seen above, for example, that $\frac{kn}{n} = \frac{k}{1} = k$ for any $n, k \in \mathbf{N}$.

(G) The definition of a fraction as a point on the number line allows us to make precise the common concept of one fraction being bigger than another. First consider the case of whole numbers. The way we put the whole numbers on the number line, a whole number m is **smaller than** another whole number n (in symbols: $\boldsymbol{m < n}$) if m is to the left of n. We expand on this fact by *defining* a fraction A to be **smaller than** another fraction B, (in symbols: $\boldsymbol{A < B}$) if A is to the left of B on the number line:



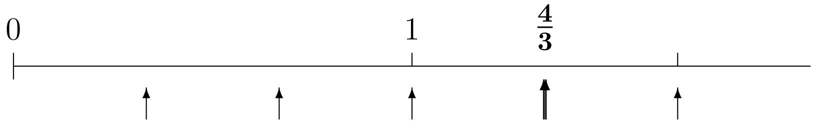
Note that in the standard education literature, the concept of $A < B$ between fractions is never defined, one reason being that if the concept of a fraction is not defined, it is difficult to say one unknown object is smaller than another unknown object.

Sometimes the symbol $\boldsymbol{B > A}$ is used in place of $A < B$. Then we say B **is bigger** than A.

This definition of *smaller than* may seem innocuous, but it is easy to overlook its significance. The concept of "bigger" or "smaller" is such a basic part of the human experience that any *definition* of either one would likely make no impression whatsoever on our psyche because we would immediately wave it off and revert to our naive conceptions. More explicitly, the inherent danger is that you would completely ignore the preceding definition of "smaller than" the next time you are called upon to decide whether $A < B$ for two fractions A and B, and would try instead to "prove" $A < B$ entirely by appealing to gut feelings. So just remember: *if you want to prove that a fraction A is smaller than a fraction B, you will have to locate the position of A and the position of B on the number line, and prove that A is to the left of B.* There is no other way.

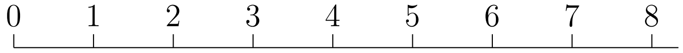
(H) There is a pedagogical issue related to the notation of a fraction: $\frac{k}{\ell}$. Students have been known to raise the question of why we use three symbols (k, ℓ, and the **fraction bar** "–") to denote one concept. With a precise definition of what a fraction is, we can easily answer this question. Remember that *a fraction is a special point on the number line*, no more and no less, and the symbols employed serve the purpose of telling us where the fraction is located. Thus the symbol $\frac{14}{5}$ says precisely that, if we look at the sequence of 5ths, then $\frac{14}{5}$ is the 14th point of the sequence to the right of 0. We need every part of the symbol $\frac{14}{5}$ for this purpose: the need of 5 and 14 is obvious, and the role of the fraction bar "–" is to separate 5 from 14 so that one does not confuse $\frac{14}{5}$ with $^{14}5$, for example.

(I) We now face the practical question of how to locate a fraction, at least approximately, on the number line. For something as simple as $\frac{4}{3}$, almost no work is involved: just divide the number line into thirds and go to the fourth point in the sequence of thirds. It is a little beyond 1.



Activity Can you locate the fraction $\frac{20}{15}$? How is it related to $\frac{4}{3}$?

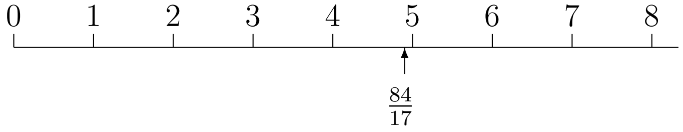
However, how to locate a fraction such as $\frac{84}{17}$, approximately, on the number line? That is, roughly, where should $\frac{84}{17}$ be placed on the following line?



Because we are trying to find out how big 84 is compared with 17, it is natural to think of division-with-remainder and divide 84 by 17. We have $84 = (4 \times 17) + 16$,[7] so

$$\frac{84}{17} = \frac{(4 \times 17) + 16}{17}$$

So if each step we take is of length $\frac{1}{17}$, going another 16 steps to the right of 4 will get us to $\frac{84}{17}$. If we go 17 steps instead, we will get to 5. Therefore $\frac{84}{17}$ should be quite near 5, as shown:



In general, if $\frac{m}{n}$ is a fraction and division-with-remainder gives $m = qn + k$, where q and k are whole numbers and $0 \leq k < n$, then

$$\frac{m}{n} = \frac{qn + k}{n} \ ,$$

and the position of $\frac{m}{n}$ on the number line would be between q $(= \frac{qn}{n}\ )$ and $q + 1$ $(=\frac{(q+1)n}{n}$, which is $\frac{qn+n}{n}$).

Decimals

There is an important class of fractions that deserves to be singled out at the outset: those fractions whose denominators are all positive powers of 10, e.g.,

$$\frac{1489}{10^2}, \quad \frac{24}{10^5}, \quad \frac{58900}{10^4}$$

[7] This is the correct way to say "84 divided by 17 has quotient 4 and remainder 16", **NOT** "$84 \div 17 = 4\ R\ 16$"! Please help to get rid of this piece of mathematical illiteracy from school classrooms.

(Recall the *exponential notation*: $10^2 = 10 \times 10$, $10^4 = 10 \times 10 \times 10 \times 10$, etc.) These are called **decimal fractions**, for obvious reasons, but they are better known in a more common notation under a slightly different name, to be described presently. Decimal fractions were understood and used in China by about 400 A.D., but they were transmitted to Europe as part of the so-called Hindu-Arabic numeral system only around the twelfth century. In 1593, the German Jesuit priest (and Vatican astronomer) C. Clavius introduced the idea[8] of writing a decimal fraction without the fraction symbol: just use the numerator and then keep track of the number of zeros in the denominator (2 in the first decimal fraction, 5 in the second, and 4 in the third of the above examples) by the use of a dot, the so-called **decimal point**, thus:

$$14.89, \quad 0.00024, \quad 5.8900,$$

respectively. The rationale of the notation is clear: the number of digits to the right of the decimal point, the so-called **decimal digits**, keeps track of the power of 10 in the respective denominators, 2 in 14.89, 5 in 0.00024, and 4 in 5.8900. In this notation, these numbers are called **finite** or **terminating decimals**. In context, we usually omit any mention of “finite” or “terminating” and just say **decimals**. Notice the convention that, in order to keep track of the power 5 in $\frac{24}{10^5}$, three zeros are added to the left of 24 to make sure that there are 5 digits to the right of the decimal point in 0.00024. Note also that the 0 in front of the decimal point is only for the purpose of clarity and is optional.

Activity Explain why $3.15 > 3.14$. (*Caution:* Remember what was said above about definitions, and be *very* careful with your explanation.)

You may be struck by the odd looking number 5.8900, because you have been told that that it is ok to omit the zeros at the right end of the decimal point and just write 5.89. But why? In other words, *why are the following two fractions equal*?

$$\frac{58900}{10^4} \quad \text{and} \quad \frac{589}{10^2}$$

They are, but your job is not finished until you can *prove* it. It is time to remember what we said earlier: *if something is asserted about fractions, then we must* prove *it*

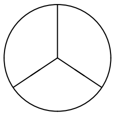
[8] See J. Ginsburg, On the early history of the decimal point, *American Mathematical Monthly*, 35 (1928), 347–349.

on the basis of the definition of a fraction as a point on the number line. Just that. Nothing more and nothing less. So why are the two points denoted by these fractions the same point on the number line? Think about it now and we will do it in the next section.

Exercises 1.1

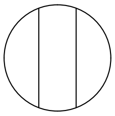
> *In doing these and subsequent exercises, please observe the following basic rules:*
>
> (*a*) ***Use only what you have learned so far*** *in this course* (*this is the situation you face when you teach*).
>
> (*b*) ***Show your work****; the explanation is as important as the answer.*
>
> (*c*) ***Be clear****. Get used to the idea that everything you say has to be understood.*

1. A text on professional development claims that students' conception of "equal parts" is fragile and is prone to errors. As an example, it says that when a circle is presented this way to students



they have no trouble shading $\frac{2}{3}$, but when these same students

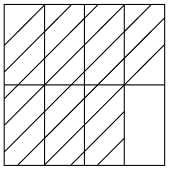
> `are asked to construct their own picture of` $\frac{2}{3}$`, we often see them create pictures with unequal pieces:`



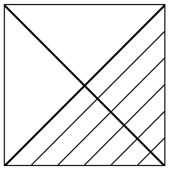
(a) What kind of faulty mathematical instruction might have promoted this kind of misunderstanding on the part of students? (b) What would you do to correct this kind of mistake by students?

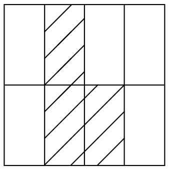
2. Indicate the approximate position of each of the following on the number line, and briefly explain why. (a) $\frac{186}{7}$. (b) $\frac{457}{13}$. (c) $\frac{39}{350}$. (d) 5.127.

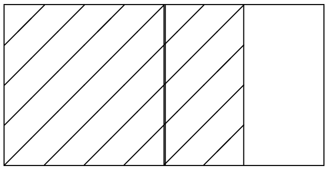
3. Suppose the unit 1 on the number line is the area of the following shaded region obtained from a division of a given square into eight congruent rectangles (and therefore eight parts of equal area).[9]



Write down the fraction of that unit representing the shaded area of each of the following divisions of the same square and *give a brief explanation of your answer.* (In the picture on the right, two copies of the same square share a common side and the square on the right is divided into two halves.)

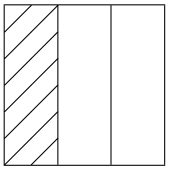






4. With the unit as in problem 3 above, write down the fraction representing the area of the following shaded region (assume that the top and bottom sides of the square are each divided into three segments of equal length):

[9] We will give a precise definition of *congruence* in Chapter 5, and will formally discuss *area* in Chapter 7. In this chapter, we only make use of both concepts in the context of triangles and rectangles, and then only in the most superficial way. For the purpose of understanding this chapter, you may therefore take both concepts in the intuitive sense. If anything more than intuitive knowledge is needed, it will be supplied on the spot, e.g., in §4 of this chapter.



5. Ellen ate $\frac{1}{3}$ of a large pizza with a 1-foot diameter and Kate ate $\frac{1}{2}$ of a small pizza with a 6-inch diameter. (Assume that all pizzas have the same thickness and that the fractions of a pizza are measured in terms of area.) Ellen told Kate that since she had eaten more pizza than Kate, $\frac{1}{3} > \frac{1}{2}$. Discuss all the mathematical mistakes in Ellen's assertion.

6. Take a pair of opposite sides of a unit square (a square whose sides have length 1) and divide each side into 478 equal parts. Join the corresponding points of division to obtain 478 thin rectangles (we will assume that these are rectangles). For the remaining pair of opposite sides, divide each into 2043 equal parts and also join the corresponding points of division; these lines are perpendicular to the other 479 lines. The intersections of these 479 and 2044 lines create 478×2043 small rectangles which are congruent to each other (we will assume that too). What is the area of each such small rectangle, *and why*? (This problem is important for §4 below.)

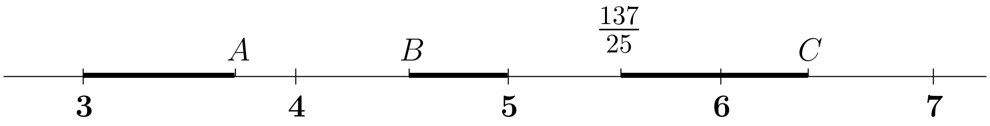
7. [*Review remark* (*B*) *in the sub-section* Some special features *on the importance of the unit before doing this problem. Also make sure that you do it by a careful use of the definition of a fraction rather than by some transcendental intuition you possess but which cannot be explained to your students.*]

(a) After driving 148 miles, we have done only two-thirds of the driving for the day. How many miles did we plan to drive for the day? Explain.

(b) After reading 180 pages of a book, I am exactly four-fifths of the way through. How many pages are in the book? Explain.

(c) Alexandra was three quarters of the way to school after having walked 0.78 mile from home. How far is her home from school? Explain.

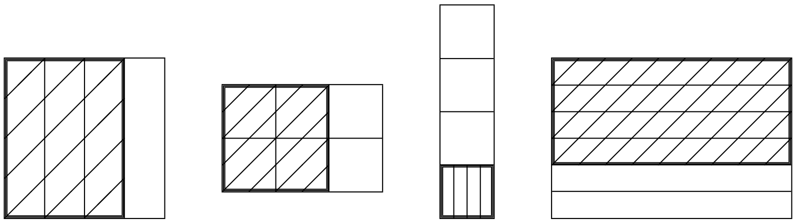
8. Three segments (thickened) are on the number line, as shown:



It is known that the length of the left segment is $\frac{11}{16}$, that of the middle segment is $\frac{8}{17}$, and that of the right segment is $\frac{23}{25}$. What are the fractions A, B, and C? (*Caution:* Remember that you have to explain your answers, and that you know nothing about "mixed numbers" until we come to this concept in §3.)

9. The following is found in a certain third-grade workbook:

> `Each of the following figures represents a fraction:`



> `Point to two figures that have the same fractions shaded.`

If you are the third grade teacher teaching from this workbook, how would you change this problem to make it suitable for classroom use?

10. A textbook for professional development defines a fraction as follows:

> `A fraction has three distinct meanings.`
>
> PART-WHOLE `The part-whole interpretation of a fraction such as` $\frac{2}{3}$ `indicates that a whole has been partitioned into three equal parts and two of those parts are being considered.`
>
> QUOTIENT `The fraction` $\frac{2}{3}$ `may also be considered as a quotient,` $2 \div 3$`. This interpretation also arises from a partitioning situation. Suppose you have some big cookies to give to three people. You could give each person one cookie, then another, and so on until you had distributed the same amount to each. If you have six cookies, then you could represent this process mathematically by` $6 \div 3$`, and each person would get two cookies. But if you only have two cookies, one way to solve the problem is to divide each cookie into three equal`

> parts and give each person $\frac{1}{3}$ of each cookie so that at the end, each person gets $\frac{1}{3}+\frac{1}{3}$ or $\frac{2}{3}$ cookies. So $2 \div 3 = \frac{2}{3}$.
>
> RATIO The fraction $\frac{2}{3}$ may also represent a ratio situation, such as there are two boys for every three girls.

Write down your reaction to this definition, including whether you believe it helps the reader see what a fraction is, and whether it makes mathematical sense. Be as precise as you can.

2 Equivalent fractions

Then fundamental theorem

How to compare fractions

The mathematics of *"of"*

Fraction as division

A little reflection

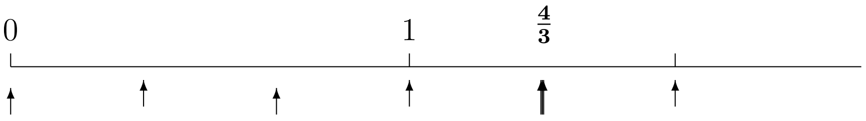
The fundamental theorem

Recall that two fractions are equal if they are the same point on the number line. We observed above that for all nonzero whole numbers n and k, $\frac{nk}{n} = \frac{k}{1}$, as both are equal to k. The following generalizes this simple fact.

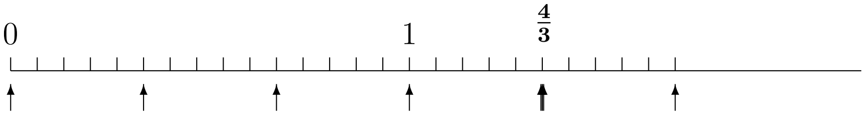
Theorem 1 *Given two fractions* $\dfrac{m}{n}$ *and* $\dfrac{k}{\ell}$*, suppose there is a nonzero whole number* c *so that* $k = cm$ *and* $\ell = cn$. *Then*

$$\frac{m}{n} = \frac{k}{\ell}$$

Proof First look at a special case: why is $\frac{4}{3}$ equal to $\frac{5\times4}{5\times3}$? We have as usual the following picture:



Now suppose we further divide each of the segments between consecutive points in the sequence of thirds into 5 equal parts. Then each of the segments $[0, 1]$, $[1, 2]$, $[2, 3]$, ... is now divided into $5 \times 3 = 15$ equal parts and, in an obvious way, we have obtained the sequence of fifteenths on the number line:



The point $\frac{4}{3}$, being the 4th point in the sequence of thirds, is now the 20th point in the sequence of fifteenths because $20 = 5 \times 4$. The latter is by definition the fraction $\frac{20}{15}$, i.e., $\frac{5\times4}{5\times3}$. Thus $\frac{4}{3} = \frac{5\times4}{5\times3}$.

The preceding reasoning is enough to prove the general case. Thus let $k = cm$ and $\ell = cn$ for whole numbers c, k, ℓ, m, and n. We will prove that $\frac{m}{n} = \frac{k}{\ell}$. In other words, we will prove:

$$\frac{m}{n} = \frac{cm}{cn}$$

The fraction $\frac{m}{n}$ is the m-th point in the sequence of n-ths. Now divide each of the segments between consecutive points in the sequence of n-ths into c equal parts. Thus each of $[0, 1]$, $[1, 2]$, $[2, 3]$, ... is now divided into cn equal parts. Thus the sequence of n-ths together with the new division points become the sequence of cn-ths. A simple reasoning shows that the m-th point in the sequence of n-ths must be the cm-th point in the sequence of cn-ths. This is another way of saying $\frac{m}{n} = \frac{cm}{cn}$. The proof is complete.

It is a tradition in school mathematics to say that two fraction (symbols) $\frac{k}{\ell}$ and $\frac{m}{n}$ are **equivalent** if they are equal, i.e., if $\frac{k}{\ell}$ and $\frac{m}{n}$ are the same point. For this reason, the content of Theorem 1 is generically called the **theorem on equivalent fractions**. Thus Theorem 1 gives a sufficient condition for two fractions $\frac{k}{\ell}$ and $\frac{m}{n}$ to be equivalent, namely, if we can find a whole number c so that $k = cm$ and $\ell = cn$.

For brevity, Theorem 1 is usually stated as

$$\frac{m}{n} = \frac{cm}{cn}$$

for all fractions $\frac{m}{n}$ and all whole numbers $c \neq 0$. In this form, Theorem 1 is called the **cancellation law** for fractions: one "cancels" the c from numerator and denominator. This is the justification for the usual method of **reducing fractions**, i.e., canceling a common divisor of the numerator and the denominator of a fraction. Thus, $\frac{51}{34} = \frac{3}{2}$ because $51 = \mathit{17} \times 3$ and $34 = \mathit{17} \times 2$. A much more substantive application of the cancellation law is to bring closure to the discussion started in §1, to the effect that the decimal 5.8900 is equal to 5.89. Recall that we had, by definition,

$$\frac{58900}{10^4} = 5.8900$$

We now show that $5.8900 = 5.89$ and, more generally, *one can add zeros to or delete zeros from the right end of the decimal point without changing the decimal.* Indeed,

$$5.8900 = \frac{58900}{10^4} = \frac{589 \times 10^2}{10^2 \times 10^2} = \frac{589}{10^2} = 5.89,$$

where the middle equality makes use of Theorem 1. The reasoning is of course valid in general, e.g.,

$$12.7 = \frac{127}{10} = \frac{127 \times 10^4}{10 \times 10^4} = \frac{1270000}{10^5} = 12.70000$$

How to compare fractions

Theorem 1 is the *fundamental fact about fractions*, and the reason can be easily seen in almost all subsequent discussions in this chapter. We give one illustration right away.

Given two fractions $\frac{k}{\ell}$ and $\frac{m}{n}$, we want to know if they are equal or if one is bigger than the other. For definiteness, consider the fractions $\frac{7}{9}$ and $\frac{4}{5}$; are they equal or not? Here we are talking about 7 copies of $\frac{1}{9}$ versus 4 copies of $\frac{1}{5}$, and the difficulty is immediately apparent: we don't know how to compare $\frac{1}{9}$ with $\frac{1}{5}$.

This is no different from asking: which is longer, 3500 yards or 3.2 km? We won't know until we find out how a yard compares with a kilometer. It is well-known that in this situation, we have to find a common unit for a yard and a kilometer; it turns out that a meter would do fine, because

1 yard = 0.9144 meters (exactly).

1 km = 1000 meters.

Therefore 3500 yards = 3200.4 meters and 3.2 km = 3200 meters. Now we can do the comparison: 3500 yards is slightly longer than 3.2 km.

To return to our initial problem, we need to find a common unit for $\frac{1}{9}$ and $\frac{1}{5}$. The theorem on equivalent fractions suggests that the fraction $\frac{1}{9\times 5}$ would serve very well as a common unit for both $\frac{1}{9}$ and $\frac{1}{5}$ because

$\frac{1}{9} = \frac{5}{9\times 5} = \frac{5}{45} = 5$ copies of $\frac{1}{45}$

$\frac{1}{5} = \frac{9}{9\times 5} = \frac{9}{45} = 9$ copies of $\frac{1}{45}$

Therefore we get:

$\frac{7}{9}$ is 7 copies of $\frac{5}{45}$, is therefore 7×5 copies of $\frac{1}{45}$, i.e., $\frac{7}{9} = \frac{35}{45}$

Similarly, $\frac{4}{5}$ is 4×9 copies of $\frac{1}{45}$, i.e., $\frac{4}{5} = \frac{36}{45}$.

It is very tempting to say at this point that since $35 < 36$, $\frac{7}{9} < \frac{4}{5}$, *but this is a non sequitur*. Why? Because to say $\frac{7}{9} < \frac{4}{5}$ is to say, by definition (!), that the point $\frac{7}{9}$ is to the left of the point $\frac{4}{5}$ on the number line. Yet nowhere in this argument is this conclusion to be found. Fortunately, our argument is already 99% complete — all we need to do is to exercise a little care by dotting the i's and crossing the t's. More precisely, if we consider the sequence of 45ths on the number line, then $\frac{7}{9}$ (which is equal to $\frac{35}{45}$) is the 35th point of the sequence, while $\frac{4}{5}$ (which is equal to $\frac{36}{45}$) is the 36th point of the sequence. Since we count the point in the sequence from left to right, starting with 0, we see that the 35th point is to the left of the 36th point, i.e., $\frac{7}{9}$ is to the left of $\frac{4}{5}$. Thus

$$\frac{7}{9} < \frac{4}{5}$$

The general case of $\frac{k}{\ell}$ and $\frac{m}{n}$ is entirely similar. Indeed, by Theorem 1,

$$\frac{k}{\ell} = \frac{nk}{n\ell} \qquad \text{and} \qquad \frac{m}{n} = \frac{\ell m}{n\ell} \tag{1}$$

We have therefore obtained new fraction symbols $\frac{nk}{n\ell}$ and $\frac{\ell m}{n\ell}$ to denote the points previously denoted by $\frac{k}{\ell}$ and $\frac{m}{n}$, respectively. In particular, the (point denoted by the) fraction $\frac{k}{\ell}$ is the nk-th point in the sequence of $n\ell$-ths while the (point denoted by the) fraction $\frac{m}{n}$ is the ℓm-th point of the sequence. Hence we have proved:

$$\text{if } \ kn = \ell m, \qquad \text{then } \ \frac{k}{\ell} = \frac{m}{n} \tag{2}$$

$$\text{if } \ kn < \ell m, \qquad \text{then } \ \frac{k}{\ell} < \frac{m}{n} \tag{3}$$

Consider assertion (2): we wish to point out that it already goes beyond Theorem 1. Let us see why this is so. Suppose we have two fractions, $\frac{m}{n}$ and $\frac{cm}{cn}$, where c is a nonzero whole number, we want to prove that $\frac{cm}{cn} = \frac{m}{n}$. Since obviously $(cm)n = (cn)m$, (2) guarantees that the fractions in question are equal, and this is exactly the conclusion of Theorem 1. However we don't want to stop here because more can be said along this line. We are going to show that the converse of each of (2) and (3) is also true. (The **converse** of an assertion "if P is true then Q is true" is the assertion "if Q is true then P is also true".) Thus we claim:

$$\text{if } \ \frac{k}{\ell} = \frac{m}{n}, \quad \text{then } \quad kn = \ell m,$$

$$\text{if } \ \frac{k}{\ell} < \frac{m}{n}, \quad \text{then } \quad kn < \ell m.$$

Let us deal with the case of equality (the first assertion), as the case of inequality is similar. If indeed if $\frac{k}{\ell} = \frac{m}{n}$, then by equation (1), we have

$$\frac{nk}{n\ell} = \frac{\ell m}{n\ell}$$

This says the nk-th point in the sequence of $n\ell$-ths is equal to the ℓm-th point of the same sequence. So $nk = \ell m$, as desired.

Let us summarize our discussion thus far. First we introduce a standard piece of mathematical terminology. A standard way to express the fact that a statement "if P is true then Q is also true" and its converse, "if Q is true then P is also true", are *both* true is to say that

P is true if and only if Q is true.

Another way to say this is:

A necessary and sufficient condition for P to be true is that Q is true.

Yet another way is:

P being true is equivalent to Q being true.

We will also use a symbolic expression for the same purpose:

$$\textbf{P is true} \iff \textbf{Q is true.}$$

With this understood, then we have proved:

Theorem 2 (Cross-Multiplication Algorithm) *Given fractions* $\frac{k}{\ell}$ *and* $\frac{m}{n}$. *Then*

(a) $\frac{k}{\ell} = \frac{m}{n}$ *if and only if* $kn \; = \; \ell m$.

(b) $\frac{k}{\ell} < \frac{m}{n}$ *if and only if* $kn \; < \; \ell m$.

This theorem, one of the most maligned in elementary school mathematics, deserves an extended commentary.

(A) We call attention to the relationship between Theorem 1 and part (a) of Theorem 2. Theorem 1 gives a *sufficient* condition for two fractions to be equal: *if* $k = cm$ *and* $\ell = cn$, *then* $\frac{k}{\ell} = \frac{m}{n}$. On the other hand, this is not a *necessary* condition, in the sense that the equality $\frac{k}{\ell} = \frac{m}{n}$ does not necessarily imply $k = cm$ and $\ell = cn$ for some whole number c. For example, Theorem 1 shows that $\frac{3}{2} = \frac{21}{14}$ (as $21 = 7 \times 3$ and $14 = 7 \times 2$), so that coupled with the previous remark about $\frac{51}{34}$, we have

$$\frac{21}{14} = \frac{51}{34}$$

In this case, there is clearly *no* whole number c so that c times 21 yields 51 and that the same c times 14 yields 34.

Part of the significance of part (a) of Theorem 2 is that it gives a correct formulation of a necessary and sufficient condition for two fractions $\frac{k}{\ell}$ and $\frac{m}{n}$ to be equal.

To continue with the previous example, what we *can* conclude from $\frac{21}{14} = \frac{51}{34}$ is that $21 \times 34 = 14 \times 51$.

Activity If two fractions $\frac{238}{153}$ and $\frac{406}{n}$ are equal, where n is a whole number. What is n?

(B) On a practical level, i.e., in terms of everyday engagement in mathematics, Theorem 2 is an indispensable tool and you should be as comfortable in using it as as you are with $3 \times 4 = 12$. In particular, it provides the only easy way to decide if two fractions are equal, e.g., $\frac{551}{247}$ and $\frac{203}{91}$ are equal because $551 \times 91 = 203 \times 247$.

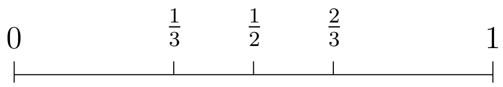
(C) The education literature often mistakes Theorem 2 for a rote-learning skill; there is as yet (2010) little or no recognition that once a fraction has been clearly defined and the equality and the concept of "less than" between two fractions has also been clearly defined, Theorem 2 has an unambiguous proof. As a result of this misunderstanding, many students have been taught to avoid using this theorem, or worse, are not taught this theorem altogether. This does gross injury to students' opportunity to learn. In these notes, we explicitly ask you to learn the proof of this theorem and make ample use of this result anytime two fractions are discussed in your teaching. We give one example:

For two nonzero whole numbers ℓ and n, $\ell > n$ is equivalent to $\frac{1}{\ell} < \frac{1}{n}$.

This is just a special case of part (b) of Theorem 2 where $k = m = 1$.

For beginners, there may be a pedagogical need to give a more transparent proof, as follows. If $\ell > n$, then $\frac{\ell}{\ell n} > \frac{n}{\ell n}$ (the ℓ-th multiple of $\frac{1}{\ell n}$ is to the right of the n-th multiple of the same), which, by the theorem on equivalent fractions (Theorem 1), is equivalent to $\frac{1}{n} > \frac{1}{\ell}$. To prove the converse, we essentially run the argument backwards: if $\frac{1}{n} > \frac{1}{\ell}$, then $\frac{\ell}{\ell n} > \frac{n}{\ell n}$, which then implies $\ell > n$.

It is common practice to dismiss the need for any *proof* of this statement, the thinking being that, for small values of ℓ and n, e.g., 3 and 2, then $3 > 2$ implies $\frac{1}{3} < \frac{1}{2}$, as the following picture clearly shows:



Therefore, the thinking is that *we can now extrapolate this intuitive argument based on picture drawing to include the general statement*, to the effect that, if $\ell > n$, then dividing $[0, 1]$ into ℓ segments of equal length would give a segment shorter than a segment obtained by dividing $[0, 1]$ into a smaller number of segments of equal length, namely, n. This kind of intuitive argument is valuable for guiding students to the correct conclusion but should not be mistaken for correct mathematical reasoning. A teacher must be always aware of the crucial distinction between the two. For example, how can this intuitive argument bring conviction to the following assertion?

$$\frac{1}{8594276} > \frac{1}{8594277}$$

This is why we should teach students the intuitive argument using small values of ℓ and n in addition to, *but not in place of,* the correct proof using the cross-multiplication algorithm. Students in sixth or seventh grade should begin to learn how to *reason mathematically* using the available facts and not rely solely on intuitive arguments.

There will be many opportunities for you to apply both parts of Theorem 2.

(D) Last but not least, we point out that, as important as Theorem 2 is, the basic idea of its proof is even more important, namely, we can regard any two fractions as two fractions with equal denominators so that their relationship can be understood at a glance. For future references, we formulate this idea as the **Fundamental Fact of Fraction-Pairs (FFFP)**:

> Any two fractions may be symbolically represented as two fractions with the same denominator.

This is no more than a restatement of equation (1). We can paraphrase FFFP this way: *any two fractions can be put on an equal footing.*

The mathematics of "of"

The power of Theorem 1 on equivalent fractions has not been fully exploited in the school mathematics literature, but it should be. We give one illustration of this fact. First, we have to give a precise meaning to a common expression, “two-thirds of something”, or more generally, “$\frac{m}{n}$ of something”. For example, what is meant by “I ate two-thirds of a pie”? Most would probably agree that this means if we look at the pie as a circular disk and cut it into 3 parts of equal area, then I ate 2 parts. Another example: what is meant by “he gave three-fifths of a bag of rice to his roommate”? Most likely, he measured his bag of rice by weight and, after dividing the bag of rice into 5 equal parts by weight, he gave away 3 parts. In each case, the choice of the unit (area in the first and weight in the second) is implicit and depends on the reader’s common sense. While the choice in each of these two cases is not controversial, one can imagine that such good fortune may not hold out in general. Consider the following example from real life[10]. A man obtained a construction loan from a bank for his house, and it stipulated that he should be at such a percentage of completion of the project at a certain point. When that time came, the bank said he had not met his obligation. Whereupon, he wrote to the bank: “Percentage of completion by what measure?” He explained that if it is computed by, say, volume of materials used, then the bank might have been correct, but if it is computed by the amount of sheer labor, then he was way ahead of schedule. The bank was flummoxed by his response.

These examples illustrate the fact that statements about “a fraction of something” could be ambiguous and, for the purpose of doing mathematics, the choice of the unit of measurement must be made explicit at the outset. It is for this reason that we are obligated to give a formal definition of “a fraction of something”. If we fix a unit of measurement, then we will use the informal language of a **quantity**, understood to be relative to the unit, to mean a number on the number line where the number 1 is the given unit.

Definition *Suppose a unit of measurement has been chosen and* $\frac{k}{\ell}$ *is a fraction. Then* $\boldsymbol{\frac{k}{\ell}}$ **of a quantity** *means the totality* (relative to this unit) *of k parts when that quantity is partitioned into ℓ equal parts according to this unit.*

The simplest quantity in the present context is that of the *length* of a segment, so that “the totality of m parts” would be the length of the concatenation of m of

[10] As related to me by my friend David Collins.

such parts. In this case, the unit of measurement will always be understood to be the length. The definition in this case then reads:

> *Let A and $\frac{k}{\ell}$ be fractions. Then $\frac{k}{\ell}$* of *A is the length of the concatenation of k parts when the segment $[0, A]$ is partitioned into ℓ parts of equal length.*

We consider some example. First, a simple one.

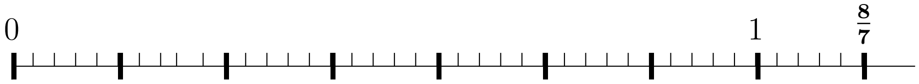
$$\tfrac{1}{3} \text{ of } \tfrac{24}{7}$$

This is then the *length* of 1 part when the segment $[0, \frac{24}{7}]$ is divided into 3 parts of equal length. Now, $\frac{24}{7}$ is 24 copies of $\frac{1}{7}$, and since $24 = 3 \times 8$, clearly $[0, \frac{24}{7}]$ can be divided into 3 equal parts so that each part is the concatenation of 8 copies of $\frac{1}{7}$. Thus $\frac{1}{3}$ of $\frac{24}{7}$ is $\frac{8}{7}$. The key point here is that the numerator of $\frac{24}{7}$ is divisible by 3. We can perhaps understand the answer a little better if we note that $\frac{8}{7} = \frac{3\times 8}{3\times 7}$, which can be written as $\frac{1\times 24}{3\times 7}$. Therefore, $\frac{1}{3}$ of $\frac{24}{7}$ is equal to $\frac{1\times 24}{3\times 7}$.

Next, suppose we want

$$\tfrac{2}{5} \text{ of } \tfrac{8}{7}$$

Now we have to divide $[0, \frac{8}{7}]$ into 5 equal parts and then measure the *length* of 2 of those parts. But first thing first: we have to divide $\frac{8}{7}$ into 5 equal parts. Noting that 8 is not divisible by 5, the key idea here is that we can make use of equivalent fractions to *force the numerator of $\frac{8}{7}$ to be divisible by 5*: we have $\frac{8}{7} = \frac{5\times 8}{5\times 7}$. The numerator 5×8 is now divisible by 5, and so by retracing the preceding steps, we conclude that if $[0, \frac{8}{7}]$ is divided into 5 equal parts, each part would be the concatenation of 8 copies of $\frac{1}{5\times 7}$. The length of two of these parts is then 2×8 copies of $\frac{1}{5\times 7}$. Thus, $\frac{2}{5}$ of $\frac{8}{7}$ is $\frac{2\times 8}{5\times 7}$. Pictorially, what we did was to sub-divide the segments between consecutive points of the sequence of sevenths into 5 equal parts:, as shown,



The unit segment is now divided into $5 \times 7 = 35$ equal parts, so that the new division points furnish *the sequence of 35ths*. The segment $[0, \frac{8}{7}]$ is now divided into 40 equal parts by this sequence of 35ths. Taking every 8th division point (in this sequence of 35ths) then gives a division of $[0, \frac{8}{7}]$ into 5 equal parts. So the length of a part in the latter division is $\frac{8}{35}$ and the length of 2 of those is of course $\frac{16}{35}$. (Of course, what we have done is merely to *re-prove the theorem on equivalent fractions in this particular case of* $\frac{8}{7} = \frac{5\times 8}{5\times 7}$*.*)

As a last example, let us compute $\frac{3}{17}$ of $\frac{11}{15}$. By Theorem 1, we have

$$\frac{11}{15} = \frac{17\times 11}{17\times 15}$$

Therefore if we divide $[0, \frac{11}{15}]$ into 17 equal parts, the length of one part is $\frac{11}{17\times 15}$, i.e., 11 copies of $\frac{1}{17\times 15}$. If want 3 parts, then it would be the concatenation of 3×11 segments of length $\frac{1}{17\times 15}$. Thus, by definition,

$$\frac{3}{17} \text{ of } \frac{11}{15} = \frac{3\times 11}{17\times 15}$$

It should be clear at this point why the following general theorem is true:

Theorem 3 *If* $\frac{k}{\ell}$ *and* $\frac{m}{n}$ *are fractions, then*

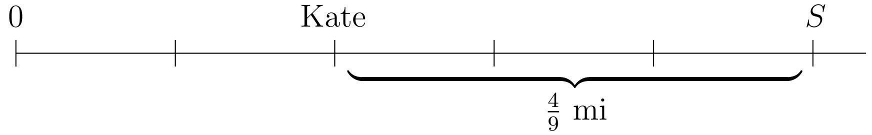
$$\frac{k}{\ell} \text{ of } \frac{m}{n} = \frac{k\,m}{\ell\,n} \tag{4}$$

Proof Because $\frac{m}{n} = \frac{\ell m}{\ell n}$, we see that $[0, \frac{m}{n}]$ is ℓm copies of $\frac{1}{\ell n}$. Therefore if we divide $[0, \frac{m}{n}]$ into ℓ equal parts, each part will be m copies of $\frac{1}{\ell n}$. Therefore if we concatenate k of these parts, we get km copies of $\frac{1}{\ell n}$, i.e., the length is $\frac{km}{\ell n}$. By the definition of $\frac{k}{\ell}$ of $\frac{m}{n}$, we have proved equation (4), and therewith also Theorem 3.

This way of exploiting equivalent fractions will be seen to clarify many aspects of fractions, such as the interpretation of a fraction as a division or the concept of multiplication (see below). It also allows us to solve word problems of the following type.

Example Kate walked $\frac{2}{5}$ of the distance from home to school, and there was still $\frac{4}{9}$ of a mile to go. How far is her home to school?

We can draw the distance from home to school on the number line, with 0 being home, the unit 1 being a mile, and S being the distance of the school from home. Then it is given that, when the segment from 0 to S is partitioned into 5 equal parts, Kate was at the second division point after 0:



If we can find the length of one of these five segments, which for convenience will be called the **short segments**, then the total distance from home to school would be 5 times that length. We are given that the distance from where Kate stands to S is $\frac{4}{9}$ of a mile, and this distance comprises 3 short segments. If we can find out how long a third of $\frac{4}{9}$ of a mile is, then we would know the length of a short segment and the problem would be solved. By the theorem on equivalent fractions,

$$\frac{4}{9} = \frac{3 \times 4}{3 \times 9} = \frac{3 \times 4}{27} \ ,$$

and this exhibits $\frac{4}{9}$ as (3×4) copies of $\frac{1}{27}$. Therefore 4 copies of $\frac{1}{27}$ (i.e., $\frac{4}{27}$) is the length of one third of $\frac{4}{9}$. The total distance from 0 to S is thus (5×4) copies of $\frac{1}{27}$, which is $\frac{20}{27}$. The distance from Kate's home to school is $\frac{20}{27}$ miles.

Remark This is one of the standard problems on fractions which is usually given after the multiplication of fractions has been introduced and the solution method is given out as an algorithm ("flip over $(1 - \frac{2}{5})$ to multiply $\frac{4}{9}$"). We see plainly that there is no need to use multiplication of fractions for the solution, and moreover, no need to memorize any solution template. The present method of solution makes the reasoning very clear.

Fraction as division

The reasoning of the preceding sub-section leads to a completely different interpretation of a fraction. We will prove:

Theorem 4 *For any fraction* $\frac{m}{n}$,

$$\frac{m}{n} \;=\; \textit{the length of one part when a segment of length } m \textit{ is partitioned into } n \textit{ equal parts}$$

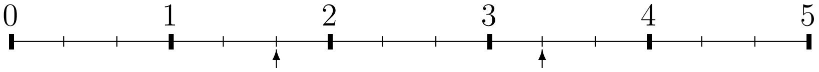
(To avoid the possibly confusing appearance of the word "divide" at this juncture, we have intentionally used "partition" instead.) Recall that the original definition of $\frac{m}{n}$ is m copies of $\frac{1}{n}$, which means to locate $\frac{m}{n}$, it suffices to consider the unit segment $[0, 1]$, divide it into n equal parts and concatenate m of these parts. The above statement, on the contrary, says that to locate $\frac{m}{n}$, one should divide, not $[0, 1]$, but $[0, m]$ into n equal parts and take the first division point to the right of 0. So the two are quite different statements.

Proof The proof is simplicity itself, but we start with a special case, say $\frac{5}{3}$, to fix our bearing. We want to prove that $\frac{5}{3}$ is the length of one part when $[0, 5]$ is divided into 3 equal parts. By the reasoning earlier in the preceding sub-section,

$$5 \;=\; \frac{5}{1} \;=\; \frac{3 \times 5}{3}$$

Thus $[0, 5]$ is the concatenation of 15 copies of $\frac{1}{3}$, and is therefore also the concatenation of 3 segments, each being the concatenation of 5 copies of $\frac{1}{3}$. Thus one of these 3 segments has length equal to 5 copies of $\frac{1}{3}$, which is $\frac{5}{3}$.

In pictures, we divide each of $[0, 1]$, . . . , $[4, 5]$ into 3 equal parts, resulting in a division of $[0, 5]$ into 3×5 equal parts. So the 5th and 10th division points give a division of $[0, 5]$ into 3 equal parts. But the 5th division point is exactly $\frac{5}{3}$ because each part is, by construction, $\frac{1}{3}$.



Now the proof in general. To divide $[0, m]$ into n equal parts, we express m as $\frac{m}{1}$, so that by Theorem 1,

$$m \;=\; \frac{nm}{n}$$

That is, $[0, m]$ is nothing but nm copies of $\frac{1}{n}$. So one part out of the n equal parts into which $[0, m]$ has been divided is just m copies of $\frac{1}{n}$, i.e., $\frac{m}{n}$. Theorem 4 is proved.

The full significance of Theorem 4 will emerge only after we re-examine the meaning of division among whole numbers. This we proceed to do. Let m, n be whole numbers and let m be a multiple of n, let us say $m = kn$ for some whole number k. Then $m \div n$ is the number of objects in a group when m objects are partitioned into n equal groups; clearly, there are k objects in each group, and therefore, $m \div n = k$. In other words, $m \div n$ is

> *the length of one part when a segment of length m is partitioned into n equal parts.*

(This is the so-called **partitive meaning** of division.) This assertion about $m \div n$ requires that m be a multiple of n at the outset, because we are doing division among whole numbers and must make sure that $m \div n$ comes out to be a whole number (i.e., the number k above). Now if we are allowed to use fractions, then for *any* m and n, the preceding indented, italicized statement (i.e., the length of one part when a segment of length m is partitioned into n equal parts) continues to make sense.[11] With this in mind, we now *define* for two arbitrary whole numbers m and n, the general concept of the **division $\boldsymbol{m \div n}$ of two whole numbers** m and n:

$\boldsymbol{m \div n}$ is the length of one part when a segment
of length $\boldsymbol{m}$ is partitioned into $\boldsymbol{n}$ equal parts.

Theorem 4 can now be rephrased as a

Theorem 5 *For* any *two whole numbers m and n, $n \neq 0$,*

$$\frac{m}{n} = m \div n$$

This is called **the division interpretation of a fraction** in school texts and in the education literature, but in that context, the subtlety of the need to define $m \div n$ when m is not a multiple of n is not clearly brought out and, moreover, such an "interpretation" is offered as another meaning of a fraction that students must memorize without benefit of explanation. We bring this fact to your attention so that, when you teach, you will make a point of giving a clear definition for $m \div n$ and also giving a clear explanation of the statement that the two numbers $\frac{m}{n}$ and $m \div n$ are equal. In mathematics, *everything* must be explained logically.

[11] In fact, m could even be a fraction.

As a result of the division interpretation of a fraction, we will retire the division symbol "$\div$" from now on and use fractions to stand for the division among whole numbers.

A little reflection

Looking back at the material of this section, you recognize that you have encountered no new facts. You have known since your school days every single statement that was put forth; in particular, you knew all of Theorems 1 through 5. So what have you learned after all?

If you are stumped by this (seemingly dumb) question, let me see if I can help you out. I would guess that whereas, before, all the things known to you were probably known to you as isolated facts, you are now made aware that they are all related to each other. For example, it may be a surprise to you that Theorem 1, the theorem on equivalent fractions — something you use to reduce fractions — is considered to be the most important single fact in fractions. Perhaps you have not given any thought to the central role it plays in the development of fractions, for example, the fact that it is the foundation that supports Theorems 2 to 5. You may have been familiar with equation (4) because you were told this is how fractions *are supposed to be* multiplied, but to be able to *prove* it, step by step, with not a trace of doubt? That is unheard of. In fact, nothing you ever read said that the phrase "$\frac{k}{\ell}$ of $\frac{m}{n}$" requires a precise definition. Isn't it enough that everybody *sort of* knows what it is? Well, actually *no*. In mathematics, if something is what everybody should know, then it must be made explicit, i.e., it must be given a *clear definition*. And so on.

But you may very well ask, "what is the point of getting to know all these connections and precise concepts?" The simple answer is that facts that hang together to tell a story are much easier to remember than unrelated, isolated ones. But there is a more substantive answer. Mathematics is, in one sense, nothing but an unending journey going from one point to the next, and the only mode of transportation is the vehicle of known facts propelled by logical reasoning. For example, one half of part (a) of Theorem 2 may be described as follows: Given

Point A: Two fractions $\frac{k}{\ell}$ and $\frac{m}{n}$ are equal.

Point B: $kn = \ell m$.

How can we go from Point A to Point B by use of only the facts known at the time of Point A plus logical reasoning? That was the problem we had to solve, and all the connections we established above were, in like manner, nothing but a living demonstration of problem-solving. Consider another example, problem 9 of Exercises 1.2 below:

> I was on a hiking trail, and after walking $\frac{7}{12}$ of a mile, I was $\frac{5}{9}$ of the way to the end. How long is the trail?

In this case, Point A and Point B are, respectively,

Point A: $\frac{5}{9}$ of a certain length of d miles is $\frac{7}{12}$ of a mile.

Point B: Exact determination of the value of d.

Again, the problem you will have to solve is how to go from Point A to Point B by use of all the facts known about fractions up to Point A plus logical reasoning. Let me emphasize: these notes or any correctly presented mathematics book will be nothing but a series of repetitions of this routine: *Going from Point A to Point B.* When we say we want students to learn to solve problems, what we mean is that they must learn how to go from Point A to Point B in the sense described.

The substantive answer to the above question is, therefore, that if we want students to be proficient at problem solving, we cannot treat mathematics as an electric circuit in which the problem-solving switch can be turned on and off at will. Students will not learn how to solve problems if we feed them isolated facts for memorization and rambling discussions in place of precise logical reasoning and yet, when the need arises, tell them to start devising strategies to go from Point A to Point B, i.e., to solve problems. Unless problem-solving is part of the daily routine of learning mathematics, students will not learn it just because we *tell them it is important.* Mathematics learning cannot be achieved by a decree.

But to solve problems, the language we use in mathematics must be clear (hence the need for precise concepts) so that Point A and Point B can be precisely identified, and the use of logical reasoning must be ever present (hence the need for constantly making connections). These notes are a *mathematics* text, and therefore necessarily engage in problem-solving at every step. The hope is that you will learn from it and engage your students in the same manner. ***We do not talk about problem-solving in these notes, for the same reason that we do not mention that these notes are written in English. There is no need. We simply solve problems from beginning to end.***

There is of course a higher level of learning mathematics, which is to make up Point A and Point B for yourself instead of being told what they are. But one thing at a time. Let us learn the basics first, and then we can aspire to the higher learning.

At this point, you probably have a better idea now about the statement made in the General Introduction, that the main goal of these notes is to "re-organize the standard materials . . . so that you will begin to recognize school mathematics as a coherent subject with every concept and skill placed in a logically correct hierarchy."

The rest of the notes will just be more of the same.

Exercises 1.2

> [Reminder] *In doing these and subsequent exercises, please observe the following basic rules:*
>
> (*a*) ***Use only what you have learned so far*** *in this course* (*this is the situation you face when you teach*).
>
> (*b*) ***Show your work****; the explanation is as important as the answer.*
>
> (*c*) ***Be clear****. Get used to the idea that everything you say has to be understood.*

1. Reduce the following fractions **to lowest terms**, i.e., until the numerator and denominator have no common divisor > 1. (You may use a four-function calculator to test the divisibility of the given numbers by various whole numbers.)

$$\frac{42}{91}, \quad \frac{52}{195}, \quad \frac{204}{85}, \quad \frac{414}{529}, \quad \frac{1197}{1273}.$$

2. Explain each of the following to an eighth-grader, *directly and without using Theorem 1 or Theorem 2,* by drawing pictures using the number line:

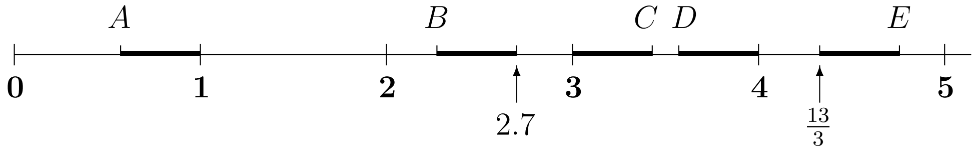
$$\frac{6}{14} = \frac{3}{7}, \quad \frac{28}{24} = \frac{7}{6}, \quad \text{and} \quad \frac{12}{27} = \frac{4}{9}\ .$$

3. School textbooks usually present the cancellation law for fractions as follows.

> Given a fraction $\frac{m}{n}$. Suppose a nonzero whole number k divides both m and n. Then $\frac{m}{n} = \frac{m \div k}{n \div k}$.

Explain as if to a seventh grader why this is true.

4. The following points on the number line have the property that the thickened segments $[A, 1]$, $[B, 2.7]$, $[3, C]$, $[D, 4]$, $[\frac{13}{3}, E]$, all have the same length:



If $A = \frac{4}{7}$, what are the values of B, C, D, E? Be careful with your explanations: we don't know how to add or subtract fractions yet. (*Rest assured that on the basis of what has been discussed in this section, you can do this problem.*)

5. (a) $\frac{7}{3}$ is $\frac{8}{11}$ of which number? (b) I was on a hiking trail, and after walking $\frac{7}{10}$ of a mile, I was $\frac{5}{9}$ of the way to the end. How long is the trail? (c) After driving 18.5 miles, I am exactly three-fifths of the way to my destination. How far away is my destination?

6. *Explain* to a sixth grade student how to do the following problem: Nine students chip in to buy a 50-pound sack of rice. They are to share the rice equally by weight. How many pounds should each person get? (If you just say, "divide 50 by 9", that won't be good enough. You must explain what is meant by "50 divided by 9", and why the answer is $5\frac{5}{9}$.)

7. (a) A wire 314 feet long is only four-fifths of the length between two posts. How far apart are the posts? (b) Helena was three quarters of the way to school after having walked $\frac{8}{9}$ miles from home. How far is her home from school?

8. (a) $\frac{3}{7}$ of a fraction is equal to $\frac{5}{6}$. What is this fraction? (b) $\frac{m}{n}$ of a fraction is equal to $\frac{k}{\ell}$. What is this fraction?

9. James gave a riddle to his friends: "I was on a hiking trail, and after walking $\frac{7}{12}$ of a mile, I was $\frac{5}{9}$ of the way to the end. How long is the trail?" Help his friends solve the riddle.

10. Prove that the following three statements are equivalent for any four whole numbers a, b, c, and d, with $b \neq 0$ and $d \neq 0$:

$$\text{(a)}\ \frac{a}{b} = \frac{c}{d}. \qquad \text{(b)}\ \frac{a}{a+b} = \frac{c}{c+d}. \qquad \text{(c)}\ \frac{a+b}{b} = \frac{c+d}{d}.$$

(One way is to prove that (a) implies (b) and (b) implies (a). Then prove (a) implies (c) and (c) implies (a).)

11. Compare the following pairs of fractions.

$$\frac{4}{9} \text{ and } \frac{3}{7}, \quad \frac{9}{29} \text{ and } \frac{4}{13}, \quad \frac{13}{17} \text{ and } 0.76, \quad \frac{12}{23} \text{ and } \frac{53}{102}.$$

(You may use a calculator to do the multiplications of the last item.)

12. Place the three fractions $\frac{13}{6}$, $\frac{11}{5}$, and $\frac{9}{4}$ on the number line and explain how they get to where they are.

13. Suppose a, b are whole numbers so that $1 < a < b$. Which is bigger: $\frac{a-1}{a}$ or $\frac{b-1}{b}$? Can you tell by inspection? What about $\frac{a+1}{a}$ and $\frac{b+1}{b}$?

14. (a) For which fraction $\frac{m}{n}$ is it true that $\frac{m}{n} = \frac{m+1}{n+1}$? (b) For which fraction $\frac{m}{n}$ is it true that $\frac{m}{n} = \frac{m+b}{n+b}$, where b is a positive whole number?

15. Prove that between $\frac{23}{123}$ and $\frac{24}{123}$, there is a fraction.

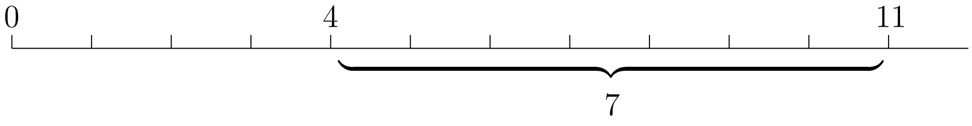
3 Adding and subtracting fractions

The meaning of adding fractions

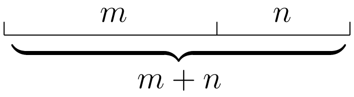
What does it mean to add $\frac{5}{7}$ to $\frac{3}{8}$?

This simple question, incredibly, is almost never answered in school mathematics. We will provide the necessary corrective measure by defining the addition of fractions as a direct extension of the addition of whole numbers. This is a point that will be stressed all through the discussion of the arithmetic operation on fractions: *They are very similar to the operations on whole numbers.* We will then *prove* the formula for adding fractions without once mentioning "the least common denominator". Contrary to what you have been told, the latter has never been an integral part of the concept of fraction addition.

Consider, for example, the addition of 4 to 7. In terms of the number line, this is just the total length of the concatenation of two segments, one of length 4 and the other of length 7, which is of course 11, as shown.



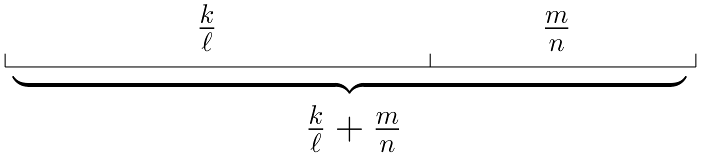
Similarly, if we have two whole numbers m and n, then $m+n$ is simply the length of the concatenation of the two segments of length m and n:



Remembering that whole numbers and fractions are on equal footing on the number line, we are therefore led to the following definition of the sum of two fractions:

Definition *Given fractions* $\frac{k}{\ell}$ *and* $\frac{m}{n}$*, we define their* **sum** *or* **addition** $\boldsymbol{\frac{k}{\ell} + \frac{m}{n}}$ *as follows:*

$$\frac{k}{\ell} + \frac{m}{n} \;=\; \textit{the length of two concatenated segments, one of length } \tfrac{k}{\ell}\textit{, followed by one of length } \tfrac{m}{n}$$



It follows directly from this definition that the addition of fractions satisfies the associative and commutative laws. (Cf. the Appendix at the end of this chapter for a summary of these laws.)

Observe that, *a priori*, the sum of two fractions may be a number that is not a fraction. However, we will remove any such suspense right away by proving that the sum of two fractions is always a fraction. First, it follows from the definition that

$$\frac{k}{\ell} + \frac{m}{\ell} \;=\; \frac{k+m}{\ell}$$

because both sides are equal to the length of $k+m$ copies of $\frac{1}{\ell}$. More explicitly, the left side is the length of k copies of $\frac{1}{\ell}$ combined with m copies of $\frac{1}{\ell}$, and is therefore the length of $k+m$ copies of $\frac{1}{\ell}$, which is exactly the right side. This tells us that, *to compute the sum of two fractions with the same denominator* ℓ*, one adds them as if they are whole numbers*, with the only difference being that instead of adding so many copies of 1, we now add so many copies of $\frac{1}{\ell}$, as above.

Because of FFFP, the general case of adding two fractions with unequal denominators is immediately reduced to the case of equal denominators, i.e., to add

$$\frac{k}{\ell} + \frac{m}{n}$$

where $\ell \neq n$, we use FFFP to rewrite $\frac{k}{\ell}$ as $\frac{kn}{\ell n}$ and $\frac{m}{n}$ as $\frac{\ell m}{\ell n}$. Then

$$\frac{k}{\ell} + \frac{m}{n} = \frac{kn}{\ell n} + \frac{\ell m}{\ell n} = \frac{kn + \ell m}{\ell n}$$

It will be observed that $\frac{kn+\ell m}{\ell n}$ is a fraction.

In a middle school classroom, it may not be a good idea to use FFFP in this peremptory fashion. One may proceed instead as follows. The reason it is not obvious how to compute the exact value of $\frac{k}{\ell} + \frac{m}{n}$ is that we are asked to add k copies of $\frac{1}{\ell}$ to m copies of $\frac{1}{n}$, which is similar to adding 5 inches to 3 meters. We cannot get an exact answer to the latter until we can express inch and meter in terms of a common unit such as centimeters, for instance. We know 1 in. = 2.54 cm. and of course 1 m. = 100 cm. Therefore

$$5 \text{ in.} + 3 \text{ m.} = (12.7 + 300) \text{ cm.} = 312.7 \text{ cm.}$$

So in the same way, we will do the addition of $\frac{k}{\ell} + \frac{m}{n}$ by first expressing both $\frac{1}{\ell}$ and $\frac{1}{n}$ in terms of a common unit, and the most obvious such unit is $\frac{1}{\ell n}$. Then $\frac{1}{\ell} = \frac{n}{\ell n}$ (which is n copies of $\frac{1}{\ell n}$), and $\frac{1}{n} = \frac{\ell}{\ell n}$ (which is ℓ copies of the same). Consequently,

$\frac{k}{\ell}$ is k copies of $\frac{1}{\ell}$, and is therefore kn copies of $\frac{1}{\ell n}$,

$\frac{m}{n}$ is m copies of $\frac{1}{n}$, and is therefore ℓm copies of $\frac{1}{\ell n}$.

Thus, $\frac{k}{\ell} + \frac{m}{n}$ is $kn+\ell m$ copies of $\frac{1}{\ell n}$, i.e., $\frac{kn+\ell m}{\ell n}$, which is the same result as before.

It is clear, by the same reasoning, that if given $\frac{m}{n}$ and $\frac{k}{\ell}$, there is a whole number D that is a common multiple of both n and ℓ, say $D = \ell_1 \ell = n_1 n$, then the computation of the sum $\frac{k}{\ell} + \frac{m}{n}$ can proceed as follows:

$$\frac{k}{\ell} + \frac{m}{n} = \frac{k\ell_1}{D} + \frac{mn_1}{D} = \frac{k\ell_1 + mn_1}{D}$$

Such a D is called a **common denominator** of the denominators n and ℓ. For example, if we are given $\frac{3}{4} + \frac{7}{8}$, then using 8 rather than 4×8 as a common denominator for the addition visibly leads to a simpler computation:

$$\frac{3}{4} + \frac{7}{8} = \frac{6}{8} + \frac{7}{8} = \frac{13}{8}$$

A more interesting example to illustrate the advantage of using a simpler common denominator can be found in problem 10 of Exercises 1.3 at the end of this section.

We add the perhaps superfluous comment that *the most obvious common denominator is the product of the denominators* (of the two fractions in question), and this is the one we use automatically.

Adding decimals

The first application of fraction addition is the explanation of the **addition algorithm for (finite) decimals**. For example, consider

$$4.0451 + 7.28$$

This algorithm calls for

(*i*) lining up 4.0451 and 7.28 by their decimal point,

(*ii*) adding the two numbers as if they are whole numbers and getting a whole number, to be called N, and

(*iii*) putting the decimal point back in N to get the answer of 4.0451+7.28.

We now supply the reasoning for the algorithm. First of all, we use FFFP. Because by definition

$$4.0451 \;=\; \frac{40451}{10^4} \quad \text{and} \quad 7.28 \;=\; \frac{728}{10^2},$$

it is obvious what common denominator to use: 10^4. So we write

$$7.28 \;=\; \frac{72800}{10^4} \;=\; 7.2800$$

We now have *two decimals with the same number of decimal digits*, i.e., $4.0451+7.28 = 4.0451 + 7.2800$. This corresponds to (*i*). Then,

$$\begin{aligned} 4.0451 + 7.28 \;&=\; \frac{40451 + 72800}{10^4} \\ &=\; \frac{113251}{10^4} && \text{(corresponds to } (ii)) \\ &=\; 11.3251 && \text{(corresponds to } (iii)) \end{aligned}$$

The reasoning is of course completely general and is applicable to any other pair of decimals.

A second application of fraction addition is to get the so-called complete expanded form of a (finite) decimal. For example, given 4.1297, we know it is the fraction

$$\frac{41297}{10^4}$$

But we have the expanded form of the whole number 41297:

$$41297 = (4 \times 10^4) + (1 \times 10^3) + (2 \times 10^2) + (9 \times 10^1) + (7 \times 10^0)$$

We also know that, by equivalent fractions, $\frac{4\times 10^4}{10^4} = 4$, $\frac{1\times 10^3}{10^4} = \frac{1}{10}$, etc. Thus

$$4.1297 \;=\; 4 + \frac{1}{10} + \frac{2}{10^2} + \frac{9}{10^3} + \frac{7}{10^4}$$

This expression of 4.1297 as a sum of descending powers[12] of 10, where the coefficients of these powers are the digits of the number itself (i.e., 4, 1, 2, 9, and 7), is called the **complete expanded form** of 4.1297. Incidentally, the latter can also be written as

$$4.1297 \;=\; 4 + 0.1 + 0.02 + 0.009 + 0.0007$$

You may have been *told* that this is true, but perhaps not the fact that it should be *proved*. In any case, here is the proof.

In the same way, a decimal $0.d_1 d_2 \cdots d_n$,[13] where each d_j is a single-digit number, has the following **complete expanded form**:

$$0.d_1 d_2 \cdots d_n \;=\; \frac{d_1}{10} + \frac{d_2}{10^2} + \cdots + \frac{d_n}{10^n}$$

Mixed numbers

With the availability of the concept of fraction addition, we can now introduce the concept of mixed numbers.[14] We have seen that, in order to locate fractions

[12] "Descending" if you think of $\frac{1}{10}$ as 10^{-1}, etc.

[13] The notation here is unfortunate: "$d_1 d_2 \cdots d_n$" is *not* the product of d_1, d_2, $\ldots$, d_n.

[14] Caution: Most textbooks introduce mixed numbers *before* defining what it means to add two fractions. Don't follow these books.

on the number line, it is an effective method to use division-with-remainder on the numerator. With the availability of the concept of addition between fractions, we are now in a position to go further than before, e.g.,

$$\frac{187}{14} = \frac{(13 \times 14) + 5}{14} = \frac{(13 \times 14)}{14} + \frac{5}{14} = 13 + \frac{5}{14}$$

Thus the sum $13 + \frac{5}{14}$, as a concatenation of two segments of lengths 13 and $\frac{5}{14}$, clearly exhibits the fraction $\frac{187}{14}$ as a point on the number line about one-third beyond the number 13. The sum $13 + \frac{5}{14}$ is usually abbreviated to $13\frac{5}{14}$ by omitting the + sign and, as such, it is called a **mixed number**. More generally, a **mixed number** is a sum $n + \frac{k}{\ell}$, where n is a whole number and $\frac{k}{\ell}$ is a *proper* fraction, and **it is usually abbreviated to just** $\boldsymbol{n\frac{k}{\ell}}$.[15] *The justification for this concept is that the whole-number part of the notation gives a clear indication of the approximate location of the fraction.*

This concept causes terror among students probably because it is usually introduced in textbooks *before* the concept of the addition of fractions is in place, with the result that deep confusion is built into the concept itself. It is for the reason of avoiding this pitfall that we have postponed the introduction of this concept until now. So just remember: a mixed number is a sum of a whole number and a proper fraction. No more, and no less.

ACTIVITY Given a mixed number $q\frac{m}{n}$, where as usual q is a whole number. Explain why $q\frac{m}{n} < q + 1$.

Final thoughts on adding fractions

Before leaving the topic of adding fractions, it is time to bring closure to the comment at the beginning of this section about the mathematical inappropriateness of the usual formula for the addition of fractions in terms of their least common denominator. Given $\frac{k}{\ell}$ and $\frac{m}{n}$, one is told how to add $\frac{k}{\ell} + \frac{m}{n}$ by first finding the lowest common denominator of the fractions, which is by definition the LCM[16] of the

[15] The discussion of fractions and decimals seems to be rife with notational problems: please note that $n\frac{k}{\ell}$ is *not* the product of n and $\frac{k}{\ell}$.

[16] Least common multiple. For a precise definition, see Chapter 3, problem 3 of Exercises 3.2.

denominators. For example, suppose we want to compute $\frac{5}{4}+\frac{7}{6}$. The LCM of 4 and 6 is 12, and $12 = \mathit{3} \times 4 = \mathit{2} \times 6$. So this method says

$$\frac{5}{4}+\frac{7}{6} \;=\; \frac{\mathit{3}\times 5}{12}+\frac{\mathit{2}\times 7}{12} \;=\; \frac{15+14}{12}=\frac{29}{12}$$

The general case is similar: if the LCM of ℓ and n is B, then letting $B = \ell'\ell = n'n$ for some whole numbers ℓ' and n', the sum of these two fractions is given as

$$\frac{k}{\ell}+\frac{m}{n} \;=\; \frac{\ell' k}{\ell'\ell}+\frac{n'm}{n'n} \;=\; \frac{\ell' k + n'm}{B}$$

In some books, this formula serves as a definition of the sum of the fractions $\frac{k}{\ell}$ and $\frac{m}{n}$. Such a definition is almost guaranteed to turn off most elementary students because it bears no resemblance to the intuitive notion of addition as "combining things". But even as a formula for addition, it is no less pedagogically objectionable because there is no reason to use the LCM of ℓ and n when the obvious multiple of both denominators, ℓn, is both adequate and natural.[17]

Please help spread the information that using the least common denominator to *define* the sum of fractions is unacceptable from every conceivable angle.

It remains to point out that we are not trying to remove the concept of least common denominator from school mathematics. *Once students have a firm mastery of the concept of addition*, if the least common denominator happens to give a shortcut in an addition problem, why not use it? However, the compulsory pursuit of the least common denominator under any circumstance is not recommended.

Subtracting fractions

We next wish to discuss the subtraction of fractions. We are handicapped by not having negative fractions at our disposal, however, so that to compute $\frac{k}{\ell}-\frac{m}{n}$, we must first make sure that $\frac{m}{n} < \frac{k}{\ell}$. Recall that the cross-multiplication algorithm (part

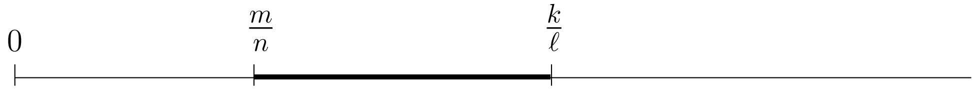
[17] *Mathematical aside:* The mathematical objection from the point of view of abstract algebra is that the requirement that we always find the LCM of the denominators is too restrictive.

(b) of Theorem 2 in §2) gives a comprehensive method to decide if one fraction is bigger than another.

The subtraction of fractions is now defined as follows: Suppose $\frac{k}{\ell} > \frac{m}{n}$, then a segment of length $\frac{k}{\ell}$ is longer than a segment of length $\frac{m}{n}$.

Definition *If* $\frac{k}{\ell} > \frac{m}{n}$, *then the* **subtraction** *or* **difference** $\boldsymbol{\frac{k}{\ell} - \frac{m}{n}}$ *is by definition the length of the remaining segment when a segment of length* $\frac{m}{n}$ *is taken from one end of a segment of length* $\frac{k}{\ell}$*.*

The segments $[0, \frac{k}{\ell}]$ and $[0, \frac{m}{n}]$ have lengths $\frac{k}{\ell}$ and $\frac{m}{n}$, of course. Therefore $\frac{k}{\ell} - \frac{m}{n}$ is just the length of the segment $[\frac{m}{n}, \frac{k}{\ell}]$:



Again as in the case of addition, the difference of two fractions is *a priori* not necessarily a fraction. However, we lay such doubts to rest immediately by showing that it is a fraction. Indeed, the same reasoning as in the case of addition, using FFFP, yields

$$\frac{k}{\ell} - \frac{m}{n} \;=\; \frac{kn - \ell m}{\ell n}$$

It is to be noted that this formula makes implicit use of the cross-multiplication algorithm in the following way: the subtraction of whole numbers in the numerator of the right side, $kn - \ell m$, does not make sense unless we know $kn > \ell m$, but this is guaranteed by the fact that $\frac{k}{\ell} > \frac{m}{n}$.

We wish to bring out the fact that *subtraction can be expressed in terms of addition.* to see this, the definition of $\frac{k}{\ell} - \frac{m}{n}$ implies that the concatenation of a segment of length $\frac{k}{\ell} - \frac{m}{n}$ and a segment of length $\frac{m}{n}$ has length $\frac{k}{\ell}$:

$$(\frac{k}{\ell} - \frac{m}{n}) + \frac{m}{n} = \frac{k}{\ell}$$

In other words, assuming $\frac{k}{\ell} > \frac{m}{n}$, then a fraction A satisfies $\frac{k}{\ell} - \frac{m}{n} = A$ if and only if it satisfies $A + \frac{m}{n} = \frac{k}{\ell}$. Thus we may regard $\frac{k}{\ell} - \frac{m}{n}$ *as the fraction* A *that satisfies the equation* $A + \frac{m}{n} = \frac{k}{\ell}$, *and this equation involves only* $+$.

Although this alternate view seems to add nothing to the concept of subtraction, the more abstract perspective does serve as a bridge to the definition of the division of fractions (see §5).

The subtraction of mixed numbers reveals a sidelight about subtraction that may not be entirely devoid of interest. Consider the subtraction of $17\frac{2}{5} - 7\frac{3}{4}$. One can do this routinely by converting the mixed numbers into fractions:

$$17\frac{2}{5} - 7\frac{3}{4} = \frac{85+2}{5} - \frac{28+3}{4} = \frac{87}{5} - \frac{31}{4} = \frac{87 \times 4 - 31 \times 5}{5 \times 4} = \frac{193}{20}.$$

However, there is another way to do the computation:

$$17\frac{2}{5} - 7\frac{3}{4} = (17 + \frac{2}{5}) - (7 + \frac{3}{4}).$$

Anticipating a reasoning that will be made routine when we come to the study of rational numbers (§3 of Chapter 2), we rewrite the right side as $(17-7) + \left(\frac{2}{5} - \frac{3}{4}\right)$. Now we are stuck because $\frac{2}{5} < \frac{3}{4}$ so that the subtraction on the right cannot be performed according to the present definition of subtraction. Using an idea that is reminiscent of the "trading" technique in the subtraction algorithm for whole numbers, we get around this difficulty by computing as follows:

$$\begin{aligned} 17\frac{2}{5} - 7\frac{3}{4} &= (16 + 1\frac{2}{5}) - (7 + \frac{3}{4}) \\ &= (16 - 7) + (1\frac{2}{5} - \frac{3}{4}) \\ &= 9 + \left(\frac{7}{5} - \frac{3}{4}\right) \\ &= 9 + \frac{13}{20} \;=\; 9\frac{13}{20} \end{aligned}$$

The whole computation looks longer than it actually is because we interrupted it with explanations. Normally, we would have done it the following way:

$$17\frac{2}{5} - 7\frac{3}{4} = (16-7) + (1\frac{2}{5} - \frac{3}{4}) = 9 + (\frac{7}{5} - \frac{3}{4}) = 9 + \frac{13}{20} = 9\frac{13}{20}$$

exactly the same as before.

Finally, there is a similar **subtraction algorithm for finite decimals** that allows finite decimals to be subtracted as if they were whole numbers provided they are aligned by the decimal points, and then the decimal point is restored at the end. The reasoning is exactly the same as the case of addition (of decimals) and will therefore be left as a problem.

Exercises 1.3

1. (a) We have an algorithm for adding two fractions: $\frac{k}{\ell} + \frac{m}{n} = \frac{kn+\ell m}{\ell n}$. Now explain as if to an eighth grader how to obtain an algorithm for adding three fractions $\frac{k}{\ell} + \frac{m}{n} + \frac{p}{q}$. Make sure you can justify the algorithm. (b) If a, b, c are nonzero whole numbers, what is $\frac{1}{ab} + \frac{1}{bc} + \frac{1}{ac}$? Simplify your answer as much as possible.

2. Show a sixth grader how to do the following problem by *using the number line*: I have two fractions whose sum is $\frac{17}{12}$ and whose *difference* (i.e., the larger one minus the smaller one) is $\frac{1}{4}$. What are the fractions? (We emphasize that no solution of simultaneous linear equations need be used. The purpose of this problem is to demonstrate the power of the number line in the teaching of school mathematics.)

3. Explain as if to a sixth grader how to get $5.09 + 7.9287 = 13.0187$.

4. Compute $78\frac{3}{54} - \frac{67}{14}$ in two different ways, and check that both give the same result. (Large numbers are used on purpose. You may use a four-function calculator to do the calculations with whole numbers, and *only* for that purpose.)

5. (a) Which is closer to $\frac{2}{7}$, $\frac{1}{3}$ or $\frac{5}{21}$? (b) Which is closer to $\frac{2}{3}$, $\frac{12}{19}$ or $\frac{9}{13}$?
(c) Which whole number is closest to the following sum?

$$\frac{12987}{13005} + \frac{114}{51}$$

(Don't forget to prove it!)

6. State the subtraction algorithm for finite decimals, and explain why it is true. (*See the discussion of the addition algorithm for finite decimals near the beginning of this*

section.)

7. (a) $\frac{2}{5} + \frac{7}{12} = ?$ (b) Laura worked on a math problem for 35 minutes without success. She came back and re-focused and got it done in 24 minutes. How much time did she spend on this problem altogether, and what does this have to do with part (a)?

8. Explain as if to a fifth grader why 1.92 is bigger than 1.91987. (*Caution:* What is a decimal?)

9. (a) We want to make some red liquid. One proposal is to mix 18 fluid ounces of liquid red dye in a pail of 230 fluid ounces of water, and the other proposal is to mix 12 fluid ounces of red dye in a smaller pail of 160 fluid ounces of water. The question: which would produce a redder liquid? *Do this problem in two different ways.* (b) An alcohol solution mixes 5 parts water with 23 parts alcohol. Then 3 parts water and 14 parts alcohol are added to the solution. Which has a higher concentration of alcohol, the old solution or the new?

10. If n is a whole number, we define $\boldsymbol{n!}$ (read: $\boldsymbol{n}$ **factorial**) to be the product of all the whole numbers from 1 through n. Thus $5! = 1 \times 2 \times 3 \times 4 \times 5$. We also define $0!$ to be 1. Define the so-called **binomial coefficients** $\left(\begin{array}{c} n \\ k \end{array}\right)$ for any whole number k satisfying $0 \le k \le n$ as

$$\left(\begin{array}{c} n \\ k \end{array}\right) = \frac{n!}{(n-k)!\; k!}$$

Then prove:

$$\left(\begin{array}{c} n \\ k \end{array}\right) = \left(\begin{array}{c} n-1 \\ k \end{array}\right) + \left(\begin{array}{c} n-1 \\ k-1 \end{array}\right)$$

(For those who remember *Pascal's triangle*, this formula describes the usual rule for generating Pascal's triangle: add two consecutive numbers in the $(n-1)$-th row to get the number right below them in the n-th row.)

11. Prove that the following statements are equivalent for fractions A, B, C, and D:

(1) $A < B \iff$ there is a fraction C so that $A + C = B$.

(2) $A < B$ implies $A + C < B + C$ for every fraction C.

(3) $A < B$ and $C < D$ implies $A + C < B + D$.

12. Let $\frac{a}{b}$ be a nonzero fraction, with $a \neq b$. Order the following (infinite number of) fractions: $\frac{a}{b}$, $\frac{a+1}{b+1}$, $\frac{a+2}{b+2}$, $\frac{a+3}{b+3}$, ... (Caution: it makes a difference whether $a < b$, or $a > b$.)

13. In the notation of problem 11, observe that each fraction $\frac{n!}{j}$, where n, j are whole numbers and $1 \leq j \leq n$, is actually a whole number. Find the following sum and simplify your answer as much as possible:

$$\frac{1}{\frac{100!}{1}} + \frac{1}{\frac{100!}{2}} + \frac{1}{\frac{100!}{3}} + \cdots + \frac{1}{\frac{100!}{98}} + \frac{1}{\frac{100!}{99}} + \frac{1}{\frac{100!}{100}}$$

14. On April 30, 2009, *Cape Cod Times* reported that in the town of Truro, MA, officials declared that voters had "narrowly approved one of four zoning amendments" by meeting the legal requirement of a two-thirds vote. It turned out that the precise vote was 136 to 70, and the officials said since the calculator gave a value of 136 to 0.66×206 when rounded to the nearest whole number, 136 was two-thirds of the total vote count of 206.

Discuss whether the town officials were right in saying 136 is two-thirds of 206 *only* using what we have learned thus far.

4 Multiplying fractions

The definition and the product formula

A mathematical comment

Area of a rectangle

Multiplication of decimals

The definition and the product formula

In the context of school mathematics, it is of vital importance that we give a definition of fraction multiplication. The reason is that this concept is one of those whose precise meaning seems to elude school textbook authors and education researchers. Recall that for whole numbers, multiplication is, by definition, just repeated addition: 3×5 means $5+5+5$ and 4×17 means $17+17+17+17$. Such a definition cannot be literally extended to fractions, e.g., it makes little sense to define $\frac{4}{7} \times \frac{1}{2}$ as "adding $\frac{1}{2}$ to itself $\frac{4}{7}$ times". Consequently, the presentation of fraction multiplication in school mathematics is usually evasive, and coincidentally, education researchers have resorted to advocating extreme measures to achieve any kind of understanding of this concept.[18]

We will do mathematics the way mathematics is normally done by giving a precise definition and drawing precise consequences. Notice once again that this definition of fraction multiplication is a direct extension of the definition of whole number multiplication.

Definition *The* **product** *or* **multiplication** *of two fractions* $\boldsymbol{\frac{k}{\ell} \times \frac{m}{n}}$ *is by definition*

the length of the concatenation of k *of the parts*
when $[0, \frac{m}{n}]$ *is partitioned into* ℓ *equal parts.*

Note that, according to the definition in §2 of "$\frac{m}{n}$ of a quantity", we may rephrase the definition as:

$$\frac{k}{\ell} \times \frac{m}{n} \;=\; \frac{k}{\ell} \;\textit{ of a segment of length }\; \frac{m}{n}$$

Or, more simply, when the unit is understood to be the unit of length:

$$\frac{k}{\ell} \times \frac{m}{n} \;=\; \frac{k}{\ell} \;\textit{ of }\; \frac{m}{n}$$

If $\ell = n = 1$, then this definition says $\frac{k}{1} \times \frac{m}{1}$ is

the length of the concatenation of k *of the parts*
when $[0, m]$ *is partitioned into 1 equal part.*

[18] For example, some suggest that one must rethink this concept by finding "multiplicative relationships between multiplicative structures" without saying what this means. *This is not helpful.*

In other words, it is the length of the concatenation of k copies of $[0, m]$. But this is exactly the usual meaning of the whole number multiplication $k \times m$, i.e., $m + m + \cdots + m$ (k times).

This definition of fraction multiplication is also consistent with everyday practice. Indeed, suppose we want to give away two-fifths of a ham (by weight), and the ham weighs $14\frac{7}{8}$ lbs. Without thinking, we would calculate the amount of ham to be given away as

$$\frac{2}{5} \times 14\frac{7}{8} \quad \text{pounds}$$

On the other hand, two-fifths of a ham by weight is (by the definition of "of" in §2) the total weight of 2 parts when we cut the ham into 5 parts of equal weight. In terms of the number line whose unit 1 represents 1 lb. of ham, the ham is represented by the number $14\frac{7}{8}$. Cutting the ham into five parts of equal weight is then the same as partitioning the segment $[0, 14\frac{7}{8}]$ into 5 segments of the same length. Therefore the weight of two-fifths of $14\frac{7}{8}$ lbs. of ham is exactly the length of 2 of these concatenated parts. Thus the above definition of fraction multiplication is a faithful, and *precise*, reflection of what is done in our daily life. At the risk of harping on the obvious, the precision comes from the precise definition of "of" given in §2.

There are two immediate consequences of the definition of fraction multiplication. The first one is a new way to look at division by a whole number. Recall that we defined $k \div \ell$ for whole numbers k and ℓ ($\ell \neq 0$) to mean the length of one part when $[0, k]$ is partitioned into ℓ equal parts. (See Theorem 4 in §2.) Thus from the definition of fraction multiplication, we have

$$k \div \ell = \frac{1}{\ell} \times k$$

More generally, we may *define*, as a direct extension of the partitive interpretation of division among whole numbers, the **division of a fraction** A **by** ℓ to be the length of one part when $[0, A]$ is partitioned into ℓ equal parts. Then it follows from the definition of fraction multiplication that $\frac{1}{\ell} \times A$ is equal to "A divided by ℓ".

Division by a whole number ℓ will henceforth be replaced by multiplication by $\frac{1}{\ell}$.

A second immediate consequence of fraction multiplication is that, if k is a whole number, then $k\times \frac{m}{n} = \frac{k}{1} \times \frac{m}{n}$, so that

$$k \times \frac{m}{n} = \underbrace{\frac{m}{n} + \cdots + \frac{m}{n}}_{k}$$

In other words, the multiplication $k\times \frac{m}{n}$ retains the meaning of repeated addition: *it is just k copies of $\frac{m}{n}$.*

It is by no means obvious from the definition of the multiplication of fractions that $\frac{k}{\ell} \times \frac{m}{n} = \frac{m}{n} \times \frac{k}{\ell}$, i.e., it is by no means clear from the definition that, for instance, $\frac{11}{7}$ of $\frac{4}{5}$ is equal to $\frac{4}{5}$ of $\frac{11}{7}$. In other words, we cannot take for granted that multiplication is commutative among fractions. However, this property follows immediately from the following *product formula*.

Theorem 1 (Product Formula) $\quad \frac{k}{\ell} \times \frac{m}{n} = \frac{km}{\ell n}$

Since $\frac{k}{\ell} \times \frac{m}{n}$ equals $\frac{k}{\ell}$ of $\frac{m}{n}$, Theorem 1 is an immediate consequence of Theorem 3 (equation (4)) in §2.

This formula, which is usually presented, in one fashion or another, as the definition of the product $\frac{k}{\ell} \times \frac{m}{n}$ in school textbooks or professional development materials, is in fact the central theorem about fraction multiplication in school mathematics. As an immediate corollary, we have:

Corollary *The multiplication of fractions is associative, commutative, and distributive.*

(See the Appendix for a summary of these laws.) We will leave the detailed proof of the Corollary to a problem at the end of the section.

As is well-known, the product formula has numerous applications. One of the simplest may be the explanation of the usual **cancellation rule** for fractions. For example, we have

$$\frac{135}{28} \times \frac{49}{9} = \frac{105}{4}$$

because we can "cancel" the 9's and 7's in the numerators and denominators. The precise reasoning is the following:

$$\begin{aligned} \frac{135}{28} \times \frac{49}{9} &= \frac{135 \times 49}{28 \times 9} && \text{(product formula)} \\ &= \frac{15 \times 9 \times 7 \times 7}{4 \times 7 \times 9} \\ &= \frac{15 \times 7}{4} && \text{(theorem on equivalent fractions)} \\ &= \frac{105}{4} \end{aligned}$$

The same reasoning of course proves that if we multiply the fractions $\frac{m\,a}{n}$ and $\frac{k}{\ell\,a}$ (where a, m, n, k ℓ are whole numbers), we can cancel the a's to get

$$\frac{m\,a}{n} \times \frac{k}{\ell\,a} = \frac{mk}{n\,\ell}$$

A mathematical comment

This definition of fraction multiplication poses a potential problem: does it make sense? Before we explain what it means, let us say right away that, while this may not be the kind of problem you want to discuss in detail in every seventh grade classroom, you might at least mention it. We could have brought it up right after the definition was given, but we didn't, because we wanted to make sure that you got the main ideas about multiplication first. Keep in mind, though, that this discussion is supposed to take place right after the definition is given, so that no result or concept that follows from the definition can be used. In particular, no product formula, because if the definition makes no sense, then the product formula won't either.

Let us illustrate what the problem is with a simple example. Consider $\frac{1}{2} \times \frac{3}{4}$. We know that $\frac{1}{2} = \frac{2}{4}$ and $\frac{3}{4} = \frac{15}{20}$. Therefore, if the definition is to make sense, we must have the equality of

$$\frac{1}{2} \times \frac{3}{4} \quad \text{and} \quad \frac{2}{4} \times \frac{15}{20},$$

where each product is computed according to the definition. This is so because

$\frac{1}{2} \times \frac{3}{4}$ is $\frac{1}{2}$ of $\frac{3}{4}$, which by Theorem 3 of §2 is equal to $\frac{3}{8}$, and

$\frac{2}{4} \times \frac{15}{20}$ is $\frac{2}{4}$ of $\frac{15}{20}$, which by Theorem 3 of §2 is equal to $\frac{30}{80}$.

And of course, $\frac{3}{8} = \frac{30}{80}$ by the theorem on equivalent fractions. Thus there is no crisis at least for this special case.

The general case turns out to be not much different. Let the following equalities between fractions be given:

$$\frac{k}{\ell} = \frac{K}{L} \quad \text{and} \quad \frac{m}{n} = \frac{M}{N}$$

Then we need to prove that, *according to the preceding definition of a product*,

$$\frac{k}{\ell} \times \frac{m}{n} = \frac{K}{L} \times \frac{M}{N}$$

In other words, we have to prove:

$$\frac{k}{\ell} \textit{ of } \frac{m}{n} = \frac{K}{L} \textit{ of } \frac{M}{N}$$

By Theorem 3 of §2, we have

$$\frac{k}{\ell} \text{ of } \frac{m}{n} = \frac{k\,m}{\ell\,n}$$

In like manner, we have

$$\frac{K}{L} \text{ of } \frac{M}{N} = \frac{K\,M}{L\,N}$$

Hence, we must prove $\frac{km}{\ell n} = \frac{KM}{LN}$. According to Theorem 2 in §2 (cross-multiplication algorithm), this would be the case if we can prove $kmLN = \ell nKM$. In other words, we have to prove:

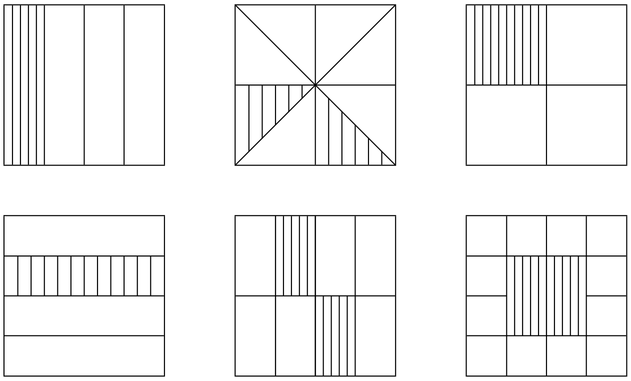
$$(kL)(mN) = (\ell K)(nM)$$

By the assumption that $\frac{k}{\ell} = \frac{K}{L}$ and by Theorem 2 of §2 again, we have $kL = \ell K$. Similarly, by the assumption that $\frac{m}{n} = \frac{M}{N}$, we also have $mN = nM$. Therefore $(kL)(mN) = (\ell K)(nM)$, as claimed.

We have therefore proved that the above definition of fraction multiplication makes sense. The mathematical terminology to express this fact is that the above definition of fraction multiplication is **well-defined**.

Area of a rectangle

A significant application of the product formula is the following well-known interpretation of fraction multiplication in terms of area;[19] this interpretation is as basic as the definition (of fraction multiplication) itself. We will prove that the area of a rectangle is equal to the product of (the lengths of) its sides. Let us first review some basic properties of area. We fix a **unit square**, i.e., a square whose sides all have length 1. The area of the unit square is **by definition** equal to 1. Congruent figures have the same area. Therefore, if the unit square is partitioned into n congruent pieces, then all these pieces have equal areas. This partition is then a division of the unit (area of the unit square) into n equal parts; by the definition of the fraction $\dfrac{1}{n}$, the area of each of these pieces is $\frac{1}{n}$. For example, each of the following shaded regions of the unit square has area equal to $\frac{1}{4}$:



For each $n = 1,\ 2,\ 3,\ 4, \ldots$, it is straightforward to get subsets of the unit square with areas equal to $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$ Any of these subsets will be referred to as **fractions of the unit square**. The area of a geometric figure which is **paved** by a finite number

[19] See footnote 8 in Exercises 1.1 after §1 concerning the concepts of *area* and *congruence*. You may take both in the naive sense in the present discussion.

of fractions of the unit square (where "pave" is used to mean the pieces overlap only at their boundaries or not at all, and their union is the figure itself) is just the sum of the areas of the latter.

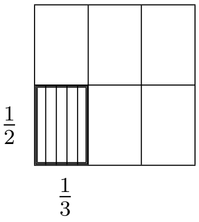
The interpretation in question is now given in the following theorem.

Theorem 2 *The area of a rectangle with sides of lengths* $\frac{k}{\ell}$ *and* $\frac{m}{n}$ *is equal to*

$$\frac{k}{\ell} \times \frac{m}{n}$$

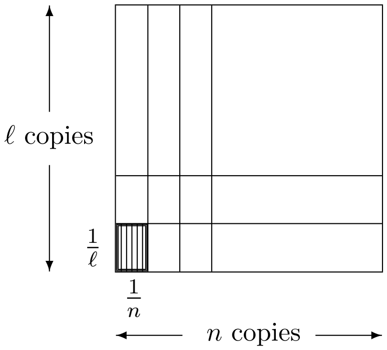
In school mathematics, this theorem has to serve as the complete statement of "area = length times width", i.e., we go only as far as fractions for the lengths of the sides of the rectangle, but nothing more. See the discussion of FASM in §7.

We first prove a simple case so as to get our bearings: the area of a rectangle whose sides have length $\frac{1}{\ell}$ and $\frac{1}{n}$ is $\frac{1}{\ell} \times \frac{1}{n}$. If $\ell = 2$ and $n = 3$, we take a unit square and divide one side into 2 halves and the other into 3 parts of equal length. Joining corresponding points of the division then partitions the square into 6 congruent rectangles:



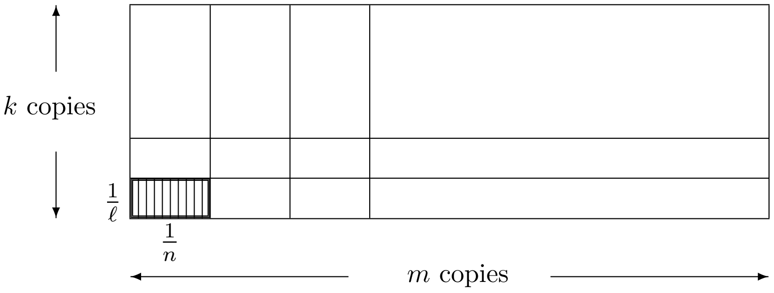
Each of the 2×3 $(= 6)$ rectangles therefore has area equal to $\frac{1}{2 \times 3}$. In particular, the shaded rectangle has sides of length $\frac{1}{2}$ and $\frac{1}{3}$, and its area is $\frac{1}{2 \times 3}$, which is equal to $\frac{1}{2} \times \frac{1}{3}$, by the product formula.

We can now give the **proof of the Theorem**. We first prove it for the special case that $k = m = 1$. Divide the two sides of a unit square into ℓ equal parts and n equal parts, respectively. Joining the corresponding division points creates a partition of the unit square into ℓn congruent rectangles (ℓ columns with n in each row), and therefore ℓn rectangles with the same area.



The area of the shaded rectangle is therefore $\frac{1}{\ell n}$, which is $\frac{1}{\ell} \times \frac{1}{n}$, by the product formula. Since $\frac{1}{\ell}$ and $\frac{1}{n}$ are the lengths of the sides of the shaded rectangle, the proof of the special case is complete.

The area of a general rectangle with sides of length $\frac{k}{\ell}$ and $\frac{m}{n}$ can now be computed. One side of such a rectangle consists of k concatenated segments each of length $\frac{1}{\ell}$, and the other consists of m concatenated segments each of length $\frac{1}{n}$. Joining corresponding division points on opposite sides leads to a partition of the big rectangle into km small rectangles each of which has sides of length $\frac{1}{\ell}$ and $\frac{1}{n}$.



We have just seen that each of these small rectangles has area equal to $\frac{1}{\ell n}$. Since the big rectangle is paved by km of these congruent small rectangles, the area of the big rectangle is the sum of the areas of these small rectangles, and is therefore equal to

$$\underbrace{\frac{1}{\ell n} + \cdots + \frac{1}{\ell n}}_{km} = \frac{km}{\ell n}$$

Thus we have proved that the area of a rectangle with sides of length $\frac{k}{\ell}$ and $\frac{m}{n}$ is $\frac{km}{\ell n}$. But by the product formula, the product $\frac{k}{\ell} \times \frac{m}{n}$ is also equal to $\frac{km}{\ell n}$. The proof of

Theorem 2 is now complete.

Multiplication of decimals

We round off the discussion of the multiplication of fractions with two remarks. The first is the explanation of the usual **multiplication algorithm for finite decimals**. Consider for example

$$1.25 \times 0.0067$$

The algorithm calls for

> (i) multiply the two numbers as if they are whole numbers by ignoring the decimal points,
>
> (ii) count the total number of decimal digits of the two decimal numbers, say p, and
>
> (iii) put the decimal point back in the whole number obtained in (i) so that it has p decimal digits.

We now justify the algorithm using this example, noting at the same time that the reasoning in the general case is the same:

$$\begin{aligned} 1.25 \times 0.0067 &= \frac{125}{10^2} \times \frac{67}{10^4} \\ &= \frac{125 \times 67}{10^2 \times 10^4} && \text{(product formula)} \\ &= \frac{8375}{10^2 \times 10^4} && \text{(corresponding to } (i)\text{)} \\ &= \frac{8375}{10^{2+4}} && \text{(corresponding to } (ii)\text{)} \\ &= 0.008375 && \text{(corresponding to } (iii)\text{)} \end{aligned}$$

A second remark is that there are two standard inequalities concerning multiplication that should be mentioned: If A, B, C, and D are fractions, then:

(A) If $A > 0$, then $AB < AC$ is equivalent to $B < C$.

(B) $A < B$ and $C < D$ imply $AC < BD$.

Both are obvious when we interpret fraction multiplication as the area of a rectangle. See problem 2 immediately following.

Exercises 1.4

1. Do each of the following without calculators. (a) $(12\frac{2}{3} \times 12\frac{2}{3} \times 12\frac{2}{3}) \times (2\frac{1}{19} \times 2\frac{1}{19} \times 2\frac{1}{19}) \times \frac{1}{26} = ?$ (b) $(\frac{7}{18} \times 4\frac{2}{3}) + (2\frac{1}{6} \times \frac{7}{18}) + (\frac{7}{18} \times 3\frac{1}{6}) = ?$ (c) $8\frac{2}{50} \times 1250\frac{1}{2} = ?$

2. Give detailed proofs of the following for fractions A, B, C, and D:

(A) If $A > 0$, then $AB < AC$ is equivalent to $B < C$.

(B) $A < B$ and $C < D$ imply $AC < BD$.

3. Give a detailed proof of the Corollary to Theorem 1.

4. Compute $2\frac{2}{5} \times 3\frac{3}{4}$ in two different ways.

5. (a) Find a fraction q so that $28\frac{1}{2} = q \times 5\frac{1}{4}$. Do the same for $218\frac{1}{7} = q \times 19\frac{1}{2}$. (b) Make up a word problem for each situation, and make sure that the problems are not the same for both.

6. The **perimeter** of a rectangle is by definition the sum of the lengths of its four sides. Show that given a fraction A and a fraction L, (a) there is a rectangle with area equal to A but with a perimeter that is bigger than L, and (b) there is a rectangle with perimeter equal to L but with an area that is less than A.

7. (a) $16\frac{1}{2}$ cups of liquid would fill a punch bowl. If the capacity of the cup is $9\frac{1}{3}$ fluid ounces, what is the capacity of the punch bowl? Explain carefully. (b) A rod can be cut into $18\frac{5}{8}$ short pieces each of which is $3\frac{1}{4}$ inches long. How long is the rod?

Explain in a way that you expect will appeal to students.

8. How many buckets of water would fill a container if the the capacity of the bucket is $3\frac{1}{3}$ gallons and that of the container is $7\frac{1}{2}$ gallons? (*Caution:* Getting an answer for this problem is easy, but *explaining* it logically is not.)

9. Give a proof of the distributive law for the division of whole numbers. Namely, let k, m, n, be whole numbers, and let $n > 0$. Then

$$(m \div n) + (k \div n) = (m + k) \div n$$

10. (*This is problem 9 in Exercises 1.2. Now do it again using the concept of fraction multiplication.*) James gave a riddle to his friends: "I was on a hiking trail, and after walking $\frac{7}{12}$ of a mile, I was $\frac{5}{9}$ of the way to the end. How long is the trail?" Help his friends solve the riddle.

11. Explain as if to a seventh grader how to multiply 2.005×0.36, and why.

12. Given two fractions. Their difference is $\frac{4}{5}$ of the smaller one, while their sum is equal to $\frac{28}{15}$. What are the fractions? (*Hint:* Use the number line.)

5 Dividing fractions

Background

The study of the concept of division among fractions begins, as usual, with the correct formulation of a definition of division. School mathematics as a whole considers the concepts of such operations as somehow *known to every student* and all that a teacher has to do is draw out this knowledge. This is one reason why these operations are never defined precisely, whether it is for whole numbers or fractions or numbers in general. This is not how mathematics is done. There must be precise definitions, and this need is nowhere greater than in the case of division.

Because whole numbers and fractions are on equal footing on the number line, we first take a broad look at division among whole numbers for guidance. We teach children that the division $\frac{36}{9}$ is equal to 4 because $4 \times 9 = 36$. This then is the statement that *36 divided by 9* is the whole number which, when multiplied by 9, gives 36. In symbols, we may express the foregoing as follows:

$\frac{36}{9}$ *is by definition the number* k *which satisfies* $k \times 9 = 36$.

Similarly, *72 divided by 24* is the whole number n which, when multiplied by 24, gives 72, i.e., $n \times 24 = 72$. Likewise, $\frac{84}{7}$ is the whole number m which satisfies $m \times 7 = 84$, etc. In general, given any two whole numbers a and b with $b \neq 0$, we always want the division $\frac{a}{b}$ to be the *whole number* c so that $cb = a$. This suggests, broadly, that the concept of division *among whole numbers* is this:

(*) *Given whole numbers* a *and* b*, then* **the division of** a **by** b*, in symbols* $\frac{a}{b}$*, is the whole number* c *so that the equality* $cb = a$ *holds.*[20]

Notice the abstract nature of this definition: we no longer say *directly* what $\frac{a}{b}$ is, but rather that it is number c that satisfies an equation $cb = a$. Compare this with the direct, and explicit, definition that $\frac{k}{\ell} + \frac{m}{n}$ is the length of the concatenation of a segment of length $\frac{k}{\ell}$ and one of length $\frac{m}{n}$. Moreover, if we use the definition in (*) as a guide for the concept of fraction division, a literal translation of (*) to fractions would give us the following tentative definition of fraction division:

(**) *Given fractions* A *and* B*, then* **the division of** A **by** B*, in symbols* $\frac{A}{B}$*, is the fraction* C *so that the equality* $A = CB$ *holds.*

[20] This precise definition of division explains why division by 0 has no meaning, because if it had, then for a nonzero whole number a, $\frac{a}{0}$ is the whole number k so that $k \times 0 = a$. But the last equation make no sense because the left side is 0 while the right side a is nonzero to start with.

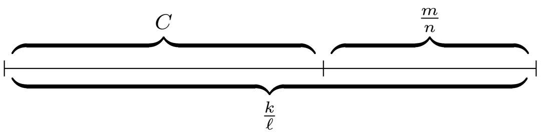
It is not usually realized, but guidance for the definition of fraction division also comes from another source: subtraction of fractions. This is because, in a naive sense, subtraction reverse what addition does, the same way that division reverses what multiplication does — at least we have seen it among whole numbers. For example, we have

$$(36 - 9) + 9 \;=\; 36,$$

and if we replace $-$ by $\div$ and $+$ by $\times$, then the preceding statement becomes the equally valid statement that

$$(36 \div 9) \times 9 \;=\; 36$$

Now consider the broader context of fractions. Here the concept of subtracting fractions is well understood, but the concept of division is yet-to-be formulated. Near the end of §3, we defined the concept of fraction subtraction and then pointed out that if fractions $\frac{k}{\ell}$ and $\frac{m}{n}$ are given, then the subtraction statement $\frac{k}{\ell} - \frac{m}{n} = C$ for some fraction C is equivalent to the addition statement $\frac{k}{\ell} = C + \frac{m}{n}$. The equivalence is immediately seen from the following picture:



In view of this equivalence, it would be legitimate to **define subtraction** of fractions in the following way:

> *Given fractions $\frac{k}{\ell}$ and $\frac{m}{n}$. Then* **the difference** $\frac{k}{\ell} - \frac{m}{n}$ *is the fraction C so that* $\frac{k}{\ell} = C + \frac{m}{n}$.

Again, notice that subtraction defined this way has an abstract flavor: we no longer say directly what the difference $\frac{k}{\ell} - \frac{m}{n}$ is, but we say, instead, that it has to be a number that satisfies an equation. Moreover, if we push the above analogy of $+$, $-$ with $\times$, $\div$, and translate this definition of subtraction literally into one about division, what we get is this:

> *Given fractions A and B. Then* **the division** $\frac{A}{B}$ *is the fraction C so that* $A = CB$.

Note that this is identical to the definition given in (**) above. This gives us confidence that the division of fractions should be defined in this manner.

Definition and invert-and-multiply

The preceding sub-section gave a freewheeling discussion that motivates how the division of fractions should be defined. Now is the time for a formal definition, and for this purpose, we have to be more careful with our language. In this and the next sub-section, we will attend to this need.

Definition *If A, B, are fractions ($B \neq 0$), then the* **division of $\boldsymbol{A}$ by $\boldsymbol{B}$**, *or the* **quotient of $\boldsymbol{A}$ by $\boldsymbol{B}$**, *denoted by* $\frac{\boldsymbol{A}}{\boldsymbol{B}}$, *is the unique fraction C so that $CB = A$.*

Because the "uniqueness" statement is used explicitly for the first time, let us first make sure that you know what it means and why it is correct. We are defining $\frac{A}{B}$ to be the fraction C so that $CB = A$, and clearly it would be a disaster if there were also another fraction C_0 so that $C_0B = A$ and $C \neq C_0$. Were this the case, what would $\frac{A}{B}$ be, C or C_0? We therefore must clear this up before going any further, and fortunately this is easy to do. For, if $CB = A = C_0B$, then $CB - C_0B = A - A = 0$. By the distributive law of fraction multiplication, we get $(C - C_0)B = 0$. We are given $B \neq 0$, so if $C - C_0 \neq 0$, then also $(C - C_0)B \neq 0$; the simplest way to see this is to use the interpretation of multiplication as area of a rectangle (Theorem 2 of §4). So we have a contradiction if we assume $C \neq C_0$. Therefore $C = C_0$, and the fraction C that satisfies $CB = A$ is unique and we are now guaranteed that $\frac{A}{B}$ has one and only one meaning.

The other possible issue with the definition is whether, with A and B given, $B \neq 0$, there *is* a fraction C so that $A = CB$. The answer is affirmative, because if $A = \frac{k}{\ell}$ and $B = \frac{m}{n}$, then the fraction $C = \frac{kn}{\ell m}$ clearly satisfies

$$A = C\,B, \quad \text{i.e.,} \quad \frac{k}{\ell} = \frac{kn}{\ell m} \times \frac{m}{n}$$

So finally, the definition of fraction division is well-defined.

For a later reference, we summarize the preceding reasoning into a theorem.

Theorem 1 *Given fractions A and B ($B \neq 0$), there is a* unique (*one and only one*) *fraction C, so that $A = CB$.*

According to the preceding discussion, if $A = \frac{k}{\ell}$ and $B = \frac{m}{n}$, then C is equal to:

$$C = \frac{k}{\ell} \times \frac{n}{m} \tag{$\natural$}$$

Therefore, according to the definition:

$$\frac{\frac{k}{\ell}}{\frac{m}{n}} = \frac{k}{\ell} \times \frac{n}{m}$$

This is the famous **invert and multiply rule** for the division of fractions. We see that there is nothing mysterious to this rule provided we make the effort to find out what it means to divide fractions. The well-known limerick, "Ours not to reason why; just invert and multiply", gets it all wrong: the problem is not the reasoning. Rather, it is a matter of whether we do mathematics according to the basic requirement of mathematics: every concept that is used must be clearly defined.

Some observations

In this sub-section, we are going to address some subtle issues surrounding the definition of division.

Despite the simplicity of the statement of Theorem 1, the theorem is conceptually sophisticated and may take some getting used to. It says, for example, that if a fraction B is nonzero, then every fraction A is a **fractional multiple** of B, in the sense that $A = CB$ for some fraction C. (Note that, since we are no longer dealing exclusively with whole numbers, the meaning of *multiple* has to be suitably modified. In the future, if we want to indicate that there is a whole number C so that $A = CB$, we will say explicitly that A is a **whole number multiple** of B.) Taking $A = 1$, the theorem implies that there is exactly one fraction, which we will denote by $\boldsymbol{B^{-1}}$, so that $B^{-1}B = 1$. We call this B^{-1} the **inverse** (or **multiplicative inverse**, to be precise) of B. If $B = \frac{m}{n}$, then the proof of the theorem shows that $B^{-1} = \frac{n}{m}$. For this reason, B^{-1} is also called the **reciprocal** of B in the context of fractions. Using this notation, the expression of C in ($\natural$) above can be rewritten as $C = AB^{-1}$. For example, if $A = \frac{11}{5}$ and $B = \frac{23}{8}$, then the C that satisfies $CB = A$ is

$$C \ = AB^{-1} = \frac{11}{5} \times \frac{8}{23} = \frac{88}{115}$$

The overriding fact concerning the concept of division is that it is defined to be is an alternate, but equivalent way of writing multiplication. This statement is more delicate than most people realize because something quite similar to it, but no longer correct, usually makes its way into most school textbooks, namely, "division and multiplication are inverse operations". Because it is a classroom issue that you must face, let us make sure you see the distinction between the two.

The statement that division is *defined* to be is an alternate, but equivalent way of writing multiplication can be made more explicit. Consider the following two statements about fractions A, B and C, with $B \neq 0$: .

(i) $\frac{A}{B} = C$

(ii) $A = CB$.

The assertion is that $(i) \iff (ii)$. Let us prove this, no matter how simple it may be. If (i) is true, then by the definition of division, C is the fraction that satisfies $A = CB$, so that (ii) is true. Conversely, suppose (ii) is true, so $A = CB$. Since $B \neq 0$, the division $\frac{A}{B}$ makes sense, and according to the definition, $\frac{A}{B}$ is the fraction so that $A \ = \frac{A}{B} \times B$. But Theorem 1 says such a fraction is unique, and since we already know $A = CB$, we conclude that $\frac{A}{B} = \ C$. This (i) is proved.

The virtue of actually writing down the proof of the equivalence is that you get to see how much it depends on having a precise definition of division and having Theorem 1 available. With this fresh in our minds, we now take a look at the usual statement in school textbooks that "multiplication and division are inverse operations". Recall that this statement is usually made at the beginning of the discussion of fraction division as a way of explaining what the latter concept means. In other words, this is the statement that serves to *define* fraction division in terms of fraction multiplication. As such, the only way it can make sense is if the concept of "inverse operation" makes sense. Now, as an operation on fractions, multiplication associates to any two fractions A and B a unique fraction AB. So *what* would an "inverse operation" be? Don't forget that division is itself an operation, and it must associate to a pair of fractions A and B a unique fraction $\frac{A}{B}$. So once again, what could be the "inverse operation" of multiplication? Is it is not clear to us here, how could it be clear to students? We must improve our textbooks.

Let us clear up another subtle point about fraction division. For a fraction such as $\frac{7}{5}$, we have explained in what sense it is a *division of 7 by 5 as whole numbers* at the end of §2. Yet $\frac{7}{5}$ may also be regarded as the division of the fraction $\frac{7}{1}$ by the fraction $\frac{5}{1}$. Are these two concepts of division the same? We now show that they are. Indeed, let us denote the division of 7 by 5 as whole numbers by the old notation $7 \div 5$ to avoid confusion. Then $7 \div 5$ is the length of one part when $[0, 7]$ is partitioned into 5 equal parts (see the end of §2). It follows that the concatenation of 5 such parts would be $[0, 7]$ and therefore $(7 \div 5) + (7 \div 5) + (7 \div 5) + (7 \div 5) + (7 \div 5) = 7$. In other words,

$$\frac{7}{1} = 5 \times (7 \div 5) = (7 \div 5) \times \frac{5}{1}$$

But according to the definition of fraction division, this says

$$\frac{\frac{7}{1}}{\frac{5}{1}} = 7 \div 5,$$

which is what we set out to prove. The same reasoning serves equally well to show that if m, n are two whole numbers ($n \neq 0$), then the meaning of m divided by n *as whole numbers* as defined at the end of §2 is the same as the meaning of $\frac{m}{1}$ divided by $\frac{n}{1}$ as fractions.

An application

The following is a typical application of the concept of fraction division in school mathematics. Notice the difference between the usual presentation in school textbooks and the one given here: *we give the explicit reason why division has to be used*, and we also *explain on purely mathematical grounds*, without resorting to any kind of bogus "conceptual understanding", the true meaning of the fractional part of the answer (i.e., the fraction $\frac{1}{40}$ below).

EXAMPLE A rod $43\frac{3}{8}$ inches long is cut into pieces which are $\frac{5}{3}$ inches long. How many such pieces can we get out of the rod?

If we change the numbers in this example to "if a rod 48 inches long is cut into pieces which are 2 inches long, how many such pieces can we get out of the rod?", then there would be no question that we do the problem by dividing 48 by 2. So we will begin the discussion by following this analogy and simply divide $43\frac{3}{8}$ by $\frac{5}{3}$ and see what we get:

$$\frac{43\frac{3}{8}}{\frac{5}{3}} = \frac{1041}{40} = 26\frac{1}{40}$$

We have used invert and multiply for the computation, of course. Now *what does the answer* $26\frac{1}{40}$ *mean?* Remembering the definition of division, we see that the preceding division is equivalent to

$$\begin{aligned} 43\frac{3}{8} &= 26\frac{1}{40} \times \frac{5}{3} \\ &= \left(26 + \frac{1}{40}\right) \times \frac{5}{3} \\ &= \left(26 \times \frac{5}{3}\right) + \left(\frac{1}{40} \times \frac{5}{3}\right) \qquad \text{(distributive law)} \end{aligned}$$

In other words, we have

$$43\frac{3}{8} = \left(26 \times \frac{5}{3}\right) + \left(\frac{1}{40} \times \frac{5}{3}\right)$$

The first term on the right, $26 \times \frac{5}{3}$, is the length of the concatenation of 26 segments each of length $\frac{5}{3}$, and the second term on the right, $\frac{1}{40} \times \frac{5}{3}$, is the length of a segment which is $\frac{1}{40}$ of $\frac{5}{3}$, by the definition of fraction multiplication. Thus the rod can be cut into 26 pieces each of $\frac{5}{3}$ inches in length, plus a piece that is only $\frac{1}{40}$ of $\frac{5}{3}$ inches. This then provides the complete answer to the problem, and retroactively justifies the use of division to do the problem.

Notice that the key to getting the correct answer is knowing the precise definition of division (which allows us to convert the division into a multiplication) and knowing the distributive law (which allows us to arrive at a correct interpretation of the answer $26\frac{1}{40}$ from the division).

You may find such an after-the-fact justification of the use of division to do the problem to be unsatisfactory. There is in fact a logical reasoning that leads inexorably to the conclusion that division should be used. We now present this reasoning.

Let there be a maximum of K copies of $\frac{5}{3}$ in $43\frac{3}{8}$, where K is a whole number. Then $43\frac{3}{8} - K \times \frac{5}{3}$ is less than $\frac{5}{3}$ (as otherwise K would not be the maximum number of such copies). Denote $43\frac{3}{8} - K \times \frac{5}{3}$ by r, then we may rewrite the definition of r as

$$43\frac{3}{8} = (K \times \frac{5}{3}) + r, \quad \textit{where } 0 \le r < \tfrac{5}{3}$$

Now, by the theorem at the beginning of this section, we may express r as a multiple of $\frac{5}{3}$, i.e., there is a fraction $\frac{m}{n}$ so that

$$r = \frac{m}{n} \times \frac{5}{3}$$

We notice that $\frac{m}{n}$ must be a **proper fraction** *in the sense that $m < n$, because $r < \frac{5}{3}$ and r is $\frac{m}{n}$ of $\frac{5}{3}$. Therefore substituting this value of r into the above equation gives:*

$$\begin{aligned} 43\frac{3}{8} &= (K \times \frac{5}{3}) + (\frac{m}{n} \times \frac{5}{3}) \\ &= (K + \frac{m}{n}) \times \frac{5}{3} \end{aligned}$$

Note that $K + \frac{m}{n}$ is a mixed number (because $\frac{m}{n}$ is a proper fraction), so we have

$$43\frac{3}{8} = (K\frac{m}{n}) \times \frac{5}{3}$$

By the definition of division, we see that

$$K\frac{m}{n} = \frac{43\frac{3}{8}}{\frac{5}{3}}$$

Of course if we know the mixed number $K\frac{m}{n}$, then we would know the answer to the problem, which is K. Therefore, the import of the preceding equation is that, in order to find the maximum number of copies of $\frac{5}{3}$ in $43\frac{3}{8}$, we should do the division:

$$\frac{43\frac{3}{8}}{\frac{5}{3}}$$

Recall that, by the above calculation, $K = 26$ and $\frac{m}{n} = \frac{1}{40}$.

We have thus explained how one can give an a priori justification for the use of division to solve this problem.

Division of finite decimals

We now bring closure to the discussion of the arithmetic of finite decimals by taking up the division of decimals. The main observation is that *the division of decimals is reduced to the division of whole numbers.* The following example is sufficient to illustrate the general case:

$$\frac{2.18}{0.625}$$

becomes, upon using invert and multiply,

$$\frac{2.180}{0.625} = \frac{\frac{2180}{10^3}}{\frac{625}{10^3}} = \frac{2180}{625}$$

This reasoning is naturally valid for the division of any two finite decimals. According to Theorem 5 of §2, the division of two whole numbers is just a fraction. Therefore the general conclusion is that *the division of any two finite decimals is equal to a fraction.*

The next step is to convert a fraction to a decimal. It turns out that in almost all cases, a fraction is equal to an infinite decimal. Referring to Theorem 2 in §2 of Chapter 3 for the precise statement, we will be content here to explain, in the special case of fractions whose denominators are a product of 2's or 5's or both, why one can convert these fractions to finite decimals by the long division of the numerator by the denominator. This is one of the most mysterious procedures in school mathematics, almost always taught by rote without any explanation. It suffices to give two examples because they already embody the general reasoning.

Consider the fraction $\frac{2180}{625}$ above. By the cancellation rule for the product of fractions (see §4), we know that for *any* whole number k,

$$\frac{2180}{625} = \left(\frac{2180 \times 10^k}{625}\right) \times \frac{1}{10^k} \qquad (\sharp)$$

Because $625 = 5^4$, the exponent 4 suggests that for $k = 4$, the fraction on the right side of ($\sharp$),

$$\left(\frac{2180 \times 10^4}{625}\right), \qquad (\flat)$$

is a whole number. The reason is that $10^4 = 2^4 \times 5^4$, so that by the cancellation law (§2),

$$\frac{2180 \times 10^4}{625} = \frac{2180 \times 2^4 \times 5^4}{5^4} = 2180 \times 2^4 = 34880$$

Therefore with $k = 4$ in ($\sharp$), we get

$$\frac{2180}{625} = 34880 \times \frac{1}{10^4} = \frac{34880}{10^4} = 3.4880$$

where the last step is by the definition of a finite decimal.

We pause to reflect on the above reasoning. First of all, the case of $k > 4$ in ($\sharp$) is immediately reduced to the case of $k = 4$, because

$$\left(\frac{2180 \times 10^k}{625}\right) \times \frac{1}{10^k} \quad = \quad \frac{2180 \times 10^4}{625} \times \frac{10^{k-4}}{10^k}$$

Now the numerator of $10^{k-4}/10^k$ is 10 times itself $k - 4$ times (don't forget $k > 4$) while the denominator is 10 times itself k times, which is 4 more 10's than what is in the numerator. Therefore

$$\frac{10^{k-4}}{10^k} = \frac{1}{10^4}$$

so that

$$\left(\frac{2180 \times 10^k}{625}\right) \times \frac{1}{10^k} \quad = \quad \left(\frac{2180 \times 10^4}{625}\right) \times \frac{1}{10^4}$$

This proves our claim.

A second comment is that if $k > 4$, the fraction in ($\flat$) with the exponent 4 replaced by k will continue to be a whole number, because

$$\frac{2180 \times 10^k}{625} = \frac{2180 \times 10^4 \times 10^{k-4}}{625} = \frac{2180 \times 10^4}{625} \times 10^{k-4} = 34880 \times 10^{k-4}$$

By ($\sharp$), we see that with any whole number $k \geq 4$, we have

$$\frac{2180}{625} = \frac{K}{10^k} \qquad (\dagger)$$

where K is the whole number

$$K = \left(\frac{2180 \times 10^k}{625}\right)$$

In practice, the explicit determination of the value of the whole number K is not done by factoring 10^4 as in the argument below (β) but *by the use of long division*, as follows. For $k = 4$, the long division of 2180×10^4 by 625 gives

$$21800000 \ = \ (34880 \times 625) + 0,$$

where the remainder is 0. Thus,

$$K \ = \ \left(\frac{34880 \times 625}{625}\right) \ = \ 34880,$$

so that, (on account of ($\sharp$)), when a decimal point is placed 4 digits from the right of 34880, we get the decimal 3.4880 that is equal to $\frac{2180}{625}$. We have thus retrieved the traditional **algorithm for converting a fraction to a decimal by long division**, at least for the special case where the denominator is a product of 2's and 5's.

We will quickly go through another example to firm up the ideas. Consider $\frac{15}{32}$. Because $32 = 2^5$, we let

$$\frac{15}{32} = \left(\frac{15 \times 10^5}{32}\right) \times \frac{1}{10^5}$$

Using long division, we obtain $1500000 = (46875 \times 32) + 0$ so that

$$\frac{15}{32} = \frac{46875}{10^5} = 0.46875$$

The fact that we know at the outset that $\frac{15 \times 10^5}{32}$ is a whole number is because

$$\frac{15 \times 10^5}{32} = \frac{15 \times 5^5 \times 2^5}{2^5} = 15 \times 5^5 = 46875$$

Moreover, as before, for any whole number $k > 5$,

$$\frac{15}{32} = \left(\frac{15 \times 10^k}{32}\right) \times \frac{1}{10^k} = \left(\frac{15 \times 10^5}{32}\right) \times \frac{1}{10^5}$$

This then leads to the usual statement that we can convert $\frac{15}{32}$ to a finite decimal by performing the long division $(15 \times 10^k) \ \div \ 32$ and then placing the decimal point k

digits from the right. The same reasoning proves the following

Theorem 2 *Let $\frac{m}{n}$ be a fraction so that n is a product of 2's and 5's. Then for a sufficiently large whole number k, the division of $m \cdot 10^k$ by n, i.e.,*

$$\frac{m \cdot 10^k}{n}$$

is a whole number q, and $\frac{m}{n}$ is equal to the finite decimal $\frac{q}{10^k}$.

To give a glimpse into the use of long division in general, consider, for example, the decimal conversion of $\frac{2}{7}$. Let us say we want 8 digits after the decimal point. Then

$$\frac{2}{7} = \frac{2 \times 10^8}{7} \times \frac{1}{10^8}$$

By the long division of 2×10^8 by 7, we get the division-with-remainder

$$2 \times 10^8 = (28571428 \times 7) + 5$$

Thus,

$$\begin{aligned} \frac{2}{7} &= \frac{(28571428 \times 7) + 5}{7} \times \frac{1}{10^8} \\ &= \frac{28571428}{10^8} + \left(\frac{5}{7} \times \frac{1}{10^8}\right) \\ &= 0.28571428 + \left(\frac{5}{7} \times \frac{1}{10^8}\right) \end{aligned}$$

What should be emphasized here is that this shows why the usual "long division of 2 by 7" (which is actually the long division of 2×10^8 by 7) yields the decimal 0.28571428, and that if we use 0.28571428 to represent $\frac{2}{7}$, then the error is at most $\frac{1}{10^8}$ (because $\frac{5}{7}$ is smaller than 1). What is left unsaid is how to make sense of this "infinite decimal" and why it is equal to the fraction $\frac{2}{7}$ itself, and why the decimal must be "repeating". Of course the "repeating" phenomenon is already making itself known through the repetition of the two-digit block "28" in 0.28571428.

Exercises 1.5

1. You want to cut pieces that are $1\frac{1}{3}$ inches long from a rod whose length is $85\frac{1}{2}$ inches. *Explain as if to a sixth grader* what is the maximum number of such pieces you can get, and how many inches of the rod are left behind.

2. It takes 2 tablespoons of a chemical to de-chlorinate 120 gallons of water. Given that 3 *tea*spoons make up a tablespoon, how many *tea*spoons of this chemical are needed to de-chlorinate x gallons of water? (Assume that the amount of water, divided by the amount of chemical needed to de-chlorinate this amount of water, is a constant.) *Caution: Don't even think about using "proportions" to do this problem.*

3. Let a, d be whole numbers, and let q and r be the quotient and remainder of a divided by d. Let also Q be the fraction so that $a = Qd$. Determine the relationship among Q, q, and r. (Those who are unsure of the meaning of division with remainder can look up §1 of Chapter 3 below.)

4. The following is an approach to the division of fractions found in some textbooks:

> `We try to find out what` $\frac{k/\ell}{m/n}$ `could mean.  Using equivalent fractions, we get`
>
> $$\frac{\frac{k}{\ell}}{\frac{m}{n}} = \frac{\frac{k}{\ell} \times \ell n}{\frac{m}{n} \times \ell n} = \frac{\frac{k\ell n}{\ell}}{\frac{m\ell n}{n}} = \frac{kn}{\ell m},$$
>
> `and therefore`
>
> $$\frac{\frac{k}{\ell}}{\frac{m}{n}} = \frac{kn}{\ell m}.$$

Is this correct?

5. The following is another approach to the division of fractions found in some textbooks:

> `We try to find out what` $\frac{k/\ell}{m/n}$ `could mean.  Using equivalent fractions, we get`
>
> $$\frac{\frac{k}{\ell}}{\frac{m}{n}} = \frac{\frac{k}{\ell} \times 1}{\frac{m}{n} \times 1} = \frac{\frac{k}{\ell} \times \frac{n}{n}}{\frac{m}{n} \times \frac{\ell}{\ell}} = \frac{\frac{kn}{\ell n}}{\frac{\ell m}{\ell n}} = \frac{kn}{\ell m},$$

`and therefore`

$$\frac{\frac{k}{\ell}}{\frac{m}{n}} = \frac{kn}{\ell m}\,.$$

Is this correct?

6. (a) How many $1\frac{1}{3}$'s are there in $95\frac{2}{7}$? (b) How many blocks of 18 minutes are there in $8\frac{1}{2}$ hours? Do it in terms of minutes, and then do it in terms of hours. Compare.

7. Prove that if B is a nonzero fraction and C is a fraction so that $CB = 0$, then $C = 0$. Find a proof that does not use Theorem 2 of §4.

8. (a) Explain as if to a sixth grader how to use long division to convert $\frac{12}{3125}$ to a decimal. (b) Do the same with $\frac{3}{64}$.

9. Do the following problem using *only* what we have done thus far: Two fractions x and y satisfy $xy = \frac{3}{10}$ and $\frac{x}{y} = \frac{8}{15}$. What are x and y?

10. (a) $\frac{5}{12}$ of a sack of rice is $8\frac{2}{3}$ the weight of 5 books. Each book weighs $2\frac{1}{2}$ lbs. How much (in lbs.) does a sack of rice weigh? (b) A pizza parlor has a Learning Fractions Special. Normally, it charges $\frac{m}{n} \times 8$ dollars for $\frac{m}{n}$ of a small pizza. During this special sale, it sells $\frac{1}{2}$ of a pizza for the price of $\frac{1}{3}$.[21] At the sales price, how much would $8\frac{2}{3}$ small pizzas cost?

11. (a) $\dfrac{1}{\frac{1}{2}(\frac{1}{3}+\frac{1}{4})} = ?$ $\dfrac{1}{\frac{1}{2}(\frac{1}{2/3}+\frac{1}{5/4})} = ?$ (b) If x, y are nonzero fractions, what is $\dfrac{1}{\frac{1}{2}(\frac{1}{x}+\frac{1}{y})}$? (This expression for x and y turns up often enough to merit a name: the **harmonic mean** of x and y.) (c) If x, y, u, v are nonzero fractions so that $x < u$ and $y < v$, prove that

$$\frac{xy}{x+y} < \frac{uv}{u+v}$$

[21] I got this idea from my friend David Collins. We believe that if all pizza parlors buy into this idea, the national fractions achievement will improve.

12. Use the number line to solve the following: If $\frac{5}{13}$ of a number N exceeds a third of N by 8, what is N ?

13. (a) Show that there is a rectangle with area < 1 sq. cm and perimeter equal to 1,000 cm. (b) Given a number A and a number L, show that there is a rectangle with perimeter equal to L cm but with an area smaller than A sq. cm.

6 Complex fractions

Why complex fractions

Basic formulas

Why complex fractions

Further applications of the concept of division cannot be given without introducing a certain *formalism* for computation about **complex fractions**, which are by definition *the fractions obtained by a division* $\frac{A}{B}$ *of two* fractions A, B ($B > 0$).[22] We continue to call A and B the **numerator** and **denominator** of $\frac{A}{B}$, respectively. Note that any complex fraction $\frac{A}{B}$ is just a fraction, more precisely, the fraction AB^{-1}, so all that we have said about fractions applies to complex fractions, e.g., if $\frac{A}{B}$ and $\frac{C}{D}$ are complex fractions, then

$$\frac{A}{B} \times \frac{C}{D} \quad \text{is} \quad \frac{A}{B} \text{ of } \frac{C}{D}.$$

Such being the case, why then do we single out complex fractions for a separate discussion? For an answer, consider a common example of adding fractions:

$$\frac{1.2}{31.5} + \frac{3.7}{0.008}$$

[22] This is a confusing piece of terminology because it suggests that *complex numbers* are involved, but they are not. Since this is the terminology in use in school mathematics and the confusion is tolerable, we will go along. Such compromises are unavoidable.

Note that this is an addition of complex fractions because $1.2 = \frac{12}{10}$, $31.5 = \frac{315}{10}$, etc. Now, the addition can be handled by the usual procedures for fractions because we may invert and multiply to obtain

$$\begin{aligned}\frac{1.2}{31.5} + \frac{3.7}{0.008} &= \frac{\frac{12}{10}}{\frac{315}{10}} + \frac{\frac{37}{10}}{\frac{8}{10^3}} \\ &= \frac{12}{315} + \frac{3700}{8} \\ &= \frac{1165596}{2520}\end{aligned}$$

Nevertheless, school students are taught to do the addition by treating the decimals *as if they were whole numbers* and directly apply the addition algorithm for fractions to get the same answer:

$$\frac{(1.2 \times 0.008) + (3.7 \times 31.5)}{31.5 \times 0.008} = \frac{116.5596}{0.252} = \frac{\frac{1165596}{10000}}{\frac{2520}{10000}} = \frac{1165596}{2520}$$

What this does is to make use of the formula $\frac{k}{\ell} + \frac{m}{n} = \frac{kn+m\ell}{\ell n}$, which is valid up to this point only for *whole numbers* k, ℓ, m, n by letting $k = 1.2$, $\ell = 31.5$, $m = 3.7$, and $n = 0.008$, regardless of the fact that 1.2, 31.5, etc., are not whole numbers. Because the simplicity of such a computation is so attractive, it gives us a strong incentive to prove that the formula

$$\frac{k}{\ell} + \frac{m}{n} = \frac{kn + m\ell}{\ell n} \quad \textit{is also valid when } k, \ell, m, n \textit{ are fractions.}$$

Similarly, we would like to be able to multiply the following *complex fractions* as if they were ordinary fractions by writing

$$\frac{0.21}{0.037} \times \frac{84.3}{2.6} = \frac{0.21 \times 84.3}{0.037 \times 2.6}$$

regardless of the fact that the product formula $\frac{k}{\ell}\frac{m}{n} = \frac{km}{\ell n}$ has only been proved for whole numbers k, ℓ, m, n.

We pause to take note of the phenomenon pointed out above, namely, the fact that school textbooks and professional development materials routinely teach skills strictly for ordinary fractions (i.e., numerator and denominator are both whole numbers) but go on to use the said skills on the more general *complex* fractions without a word of explanation. In one sense, no harm is done because, in view of items (a) to (e) below, such skills are indeed applicable to complex fractions. In a real sense though, the damage such practices do to mathematics learning may be irreparable. Why? Because such practices send multiple messages to students, among them we can list the following two: (a) *If their teachers can apply skills, without comment, to situations beyond those where the skills are supposed to be applicable, students feel that they are free to do likewise. Therefore they can make things up as they go along.* (b) *Since their teachers do not mean what they say, they consider it likely that all of mathematics is like that too. Thus following logical arguments closely and precisely ceases to be a good idea; they are forced to conclude that they must improvise at will in order to survive in the mathematical jungle.* If either idea gets to a student's head, we can forget mathematics learning.

So far we have talked about perhaps nothing more than a subjective preference for formal simplicity in calculations with complex fractions. However, the need for extending the usual formulas for ordinary fractions to complex fractions is real. For example, suppose we consider the multiplication of so-called rational expressions in a number x, e.g.,

$$\frac{x+1}{x^2-5} \cdot \frac{7}{x^3+2} = \frac{(x+1)\cdot 7}{(x^2-5)(x^3+2)}$$

This x can take any value; in particular, suppose $x = \frac{3}{4}$. Then the left side becomes a product of complex fractions:

$$\frac{\frac{3}{4}+1}{(\frac{3}{4})^2-5} \times \frac{7}{(\frac{3}{4})^3+2}$$

The fact that this product is equal to the right side, i.e., equal to

$$\frac{(\frac{3}{4}+1)\times 7}{((\frac{3}{4})^2-5)((\frac{3}{4})^3+2)}$$

therefore depends on the fact that the product formula $\frac{k}{\ell}\frac{m}{n} = \frac{km}{\ell n}$ is valid for complex fractions, i.e., for *fractions* k, ℓ, m, n. Similar computations with rational expressions related to addition or subtraction abound in the study of algebra. In other words, the validity of the usual algebraic computations *requires* an extension of the usual formalism in fractions to complex fractions. The need of coming to terms with complex fractions for the sake of learning algebra, and hence for the sake of learning middle school mathematics, is thus manifest.

It is considerations of this type that force us to take a serious look at complex fractions.

Basic formulas

Here is a brief summary of the basic facts about complex fractions that figure prominently in school mathematics: Let A, ..., F be *fractions*, and we assume further that they are nonzero where appropriate in the following. Then:

(a) Cancellation law: If $\boldsymbol{C \neq 0}$, then $\boldsymbol{\frac{AC}{BC} = \frac{A}{B}}$.

Example: $\dfrac{\frac{16}{5} \times \frac{7}{17}}{\frac{2}{3} \times \frac{7}{17}} = \dfrac{\frac{16}{5}}{\frac{2}{3}}$.

(b) $\boldsymbol{\frac{A}{B} = \frac{C}{D}}$ if and only if $\boldsymbol{AD = BC}$.

$\boldsymbol{\frac{A}{B} < \frac{C}{D}}$ if and only if $\boldsymbol{AD < BC}$

Example: $\dfrac{\frac{4}{5}}{\frac{2}{3}} < \dfrac{\frac{13}{2}}{\frac{16}{3}}$ because $\dfrac{4}{5} \times \dfrac{16}{3} < \dfrac{2}{3} \times \dfrac{13}{2}$.

(c) $\boldsymbol{\frac{A}{B} \pm \frac{C}{D} = \frac{(AD)\pm(BC)}{BD}}$

Example: $\dfrac{1.2}{31.5} + \dfrac{3.7}{0.008} = \dfrac{(1.2 \times 0.008) + (31.5 \times 3.7)}{31.5 \times 0.008}$.

(d) $\boldsymbol{\frac{A}{B} \times \frac{C}{D} = \frac{AC}{BD}}$

Example: $\dfrac{0.21}{0.037} \times \dfrac{84.3}{2.6} = \dfrac{0.21 \times 84.3}{0.037 \times 2.6}$.

(e) Distributive law: $\boldsymbol{\frac{A}{B} \times \left(\frac{C}{D} \pm \frac{E}{F}\right) = \left(\frac{A}{B} \times \frac{C}{D}\right) \pm \left(\frac{A}{B} \times \frac{E}{F}\right)}$

$$\text{Example: } \frac{0.5}{1.7} \times \left(\frac{\frac{2}{3}}{\frac{4}{5}} + \frac{\frac{6}{7}}{\frac{8}{9}} \right) \;=\; \left(\frac{0.5}{1.7} \times \frac{\frac{2}{3}}{\frac{4}{5}} \right) + \left(\frac{0.5}{1.7} \times \frac{\frac{6}{7}}{\frac{8}{9}} \right).$$

Formulas (a), (b), and (d) are the generalized versions of the cancellation law, the cross-multiplication algorithm, and the product formula, respectively, for ordinary fractions. Formula (e) is nothing more than the usual distributive law stated in the context of complex fractions, as each of $\frac{A}{B}$, $\frac{C}{D}$, etc., is just a fraction. We call explicit attention to the fact that (c) and (d) justify the above computations with $\frac{1.2}{31.5} + \frac{3.7}{0.008}$ and $\frac{0.21}{0.037} \times \frac{84.3}{2.6}$. Note also that it follows immediately from (a) that the cancellation rule for fractions (see §4) continues to hold for complex fractions: $\frac{CE}{D} \times \frac{A}{BE} = \frac{AC}{BD}$, if $E \neq 0$. For example,

$$8.7 \times \frac{125}{26.1} = 8.7 \times \frac{125}{3 \times 8.7} = \frac{125}{3}$$

One can give algebraic proofs of (a)–(d) that are entirely mechanical: e.g., for (a), let $A = \frac{k}{\ell}$, $B = \frac{m}{n}$, $C = \frac{p}{q}$, substitute these values into both sides of (a), invert and multiply each side separately and verify that the two sides are equal. Do the same for every other assertion. This way of proving (a)–(d) would be correct, but it would also not be particularly educational. We now explain a more sophisticated method of proving (a)–(e); it is one that you would use in a school classroom perhaps only sparingly, but it is a piece of mathematics that is worth learning.

Let us prove (a), i.e., $\frac{AC}{BC} = \frac{A}{B}$. Let $x = \frac{AC}{BC}$ and $y = \frac{A}{B}$. We have to prove $x = y$. Since $x = \frac{AC}{BC}$, by the definition of division, $AC = xBC$. Similarly, $A = yB$, so that multiplying both sides by C gives $AC = yBC$. Comparing this with $AC = xBC$, we see that we have expressed AC as a multiple of BC in two ways. Since $BC \neq 0$ (it is the denominator of $\frac{AC}{BC}$), Theorem 1 in §5 says that these two ways are the same, i.e., $x = y$.

The proofs of the others can be safely left as exercises.

There will be no end of examples to illustrate the ubiquity of these formulas in subsequent computations, but we can give an interesting application right away.

Example 1 Give the approximate location of $\frac{82}{26\frac{1}{2}}$ on the number line.

What we want to say, intuitively, is that $26\frac{1}{2}$ is more or less 26, and therefore the given complex fraction is roughly $\frac{82}{26}$, which is $3\frac{4}{26}$, which is a little beyond 3 on the number line. Here is one way to convert such intuitive feelings into solid mathematics. (The ability to do such conversions is a basic part of mathematics learning.) We wish to compare this clumsy complex fraction with an ordinary fraction, and there is no better way to do that than replacing $26\frac{1}{2}$ with the whole numbers closest to it: 26 and 27. Clearly, $26 < 26\frac{1}{2} < 27$, and since (intuitively) the smaller the denominator, the bigger the fraction if the numerator is fixed, we expect

$$\frac{82}{27} < \frac{82}{26\frac{1}{2}} < \frac{82}{26}$$

Having made this guess, we must prove it. Let us first prove the left inequality. By (b) above, the inequality

$$\frac{82}{27} < \frac{82}{26\frac{1}{2}}$$

is equivalent to $82 \times 26\frac{1}{2} < 82 \times 27$, and this is true because $26\frac{1}{2} < 27$ and because of the fact that $A < B$ implies $AC < BC$ (see the end of §4). One proves in a similar manner the other inequality:

$$\frac{82}{26\frac{1}{2}} < \frac{82}{26}$$

Thus the given complex fraction is trapped between $\frac{82}{27}$ and $\frac{82}{26}$, i.e., between $3\frac{1}{27}$ and $3\frac{4}{26}$. Since both of the latter are less than $3\frac{4}{24} = 3\frac{1}{6}$, the given complex fraction is beyond 3 but to the left of $3\frac{1}{6}$ on the number line.

Exercises 1.6

1. Prove (b)–(d), not by the mechanical procedure, but by employing the reasoning used in the text to prove (a).

2. Explain, in as simple a manner as possible, approximately where the fraction $\dfrac{163\frac{2}{65}}{54\frac{1}{27}}$ is on the number line. (This is a *mathematical* problem, which means that you have to be precise even when you make approximations. If you need a model, look at Example 1 above and learn how to give both an upper and lower bound for

each approximation.)

3. Let A and B be fractions and $B \neq 0$. Prove that for any nonzero whole number j,

$$\underbrace{\frac{A}{B} + \cdots + \frac{A}{B}}_{j} = \frac{jA}{B}.$$

4. Divide 98 into two parts A and B (i.e., $A + B = 98$) so that $\frac{A}{B} = \frac{6}{7}$.

5. Divide $\frac{2}{7}$ into two parts A and B so that $\frac{A}{B} = \frac{4}{5}$.

7 FASM

In this section, we give a brief indication of the role of fractions, or more generally, rational numbers (positive and negative fractions), in school mathematics. It is an informal discussion and if some statements escape you for the moment, you should just forge ahead and return to them later if necessary.

These notes only treat rational numbers. **Real numbers**, i.e., all the points on the number line, are strictly the purview of college mathematics. For example, consider the following operation with real numbers:

$$\frac{2}{\sqrt{3}} + \frac{\sqrt{5}}{4} = \frac{(4 \times 2) + (\sqrt{3}\,\sqrt{5})}{4\sqrt{3}}$$

In school mathematics, one does not explain what $\frac{2}{\sqrt{3}}$ and $\frac{\sqrt{5}}{4}$ are, much less the meaning of adding the numbers on the left. By the same token, the meaning of the product $\sqrt{3}\sqrt{5}$ on the right is even more of a mystery. In school mathematics, this difficulty has never been confronted honestly. *Implicitly*, however, the way school mathematics deals with such arithmetic operations is to appeal to what we call the **Fundamental Assumption of School Mathematics (FASM)**, which states that

> ***if an identity or an inequality "$\le$" among numbers is valid for all fractions (respectively, all rational numbers), then it is***

also valid for all nonnegative real numbers (***respectively, all real numbers.***)[23]

The validity of FASM is beyond doubt, but its proof involves considerations of limits. FASM will be a dominant theme throughout these notes. For example, the above equality is clearly patterned after the following addition formula first proved for fractions in §3:

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} \tag{5}$$

In §6, we extended this formula to allow a, b, c, d to be fractions, and in §5 of Chapter 2, we will prove that the same equality also holds for all rational numbers a, b, c, d ($bd \neq 0$), even though this is rarely done in standard texts. Therefore, by FASM, the same equality is valid for all *real* numbers a, b, c, d ($bd \neq 0$), rational or irrational. This is how we could let $a = 2$, $b = \sqrt{3}$, $c = \sqrt{5}$, and $d = 4$ to get the previous result, regardless of the fact that $\sqrt{3}$ and $\sqrt{5}$ are not rational. Clearly, FASM makes it mandatory that every school teacher, regardless of grade level, acquire a firm grasp of fractions and rational numbers.

We note in passing that this equality reveals why it is important to have a general formula for the addition of two fractions (as in (5) above), and *why the common way to* define *the addition of fractions by seeking the least common denominator distorts what fraction addition means.*

As a result of FASM, *we can now extend the definition of the division of fractions* (in §5) *to include the division of any two real numbers.* More precisely, FASM yields a version of Theorem 1 in §5 for real numbers A and B, so that given two real numbers A and B (rational or irrational), the *division of A by B*, denoted by $\frac{A}{B}$ (assuming $B \neq 0$), is now by definition the real number so that $A = (\frac{A}{B})B$. What is important for school mathematics is the fact that, *on a formal level*, FASM together with the formulas of §6 allow us to treat the division of real numbers *operationally* as the division of two fractions. Therefore, the division of real numbers can hardly be simpler from a computational point of view. With this understood, we are now in a position to take up the concepts of ratio and rate in the next section.

[23] You will not see any mention of FASM in the school mathematics literature.

8 Percent, ratio, and rate problems

Percent

Ratio

Constant rate

Units and dimension analysis

Cooperative work

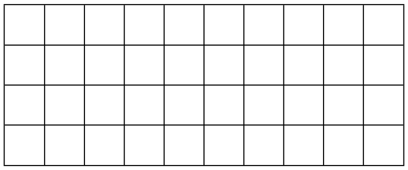
The concepts of percent, ratio, and rate are among the most troublesome for students. Concerning this phenomenon, what we know with certainty is that these concepts have never been clearly explained in the mathematics education literature. If something is never adequately taught, then it is difficult to even make a guess about the root cause of the associated learning difficulty. Any hope of improvement therefore must begin with a mathematically adequate presentation of the material in the school classroom.

In this section, we supply precise definitions of the first two by making essential use of complex fractions. We also explain why there is no need for any definition of the third (rate), but that the concept of complex fractions makes possible a lucid discussion of so-called "rate problems".

Percent

A teacher hands out the following problem to his seventh grade class:[24]

`Shade 6 of the small squares in the rectangle shown below.`



[24] M.K. Stein, M.S. Smith,, M.A. Henningsen, E.A. Silver, *Implementing Standards-Based Mathematics Instruction*, Teachers College, Columbia University, 2000. P. 47.

```
Using this diagram, explain how to determine the percent of the area
that is shaded.
```

His goal was for the students to figure out the percent representation of shaded portions of a series of rectangles. In particular, he wanted his students to "use the visual diagrams to determine their numerical answers rather than relying on the traditional algorithms" that students had learned. He was hoping that this would help students develop "conceptual understandings of [this form] of representing fractional quantities…". It turned out that, after 30 minutes, his students had no success.

We will come back to this problem after we have given a precise definition of percent in terms of complex fractions.

Definition *A* **percent** *is a complex fraction whose denominator is 100.*

The importance for a student to be able to think of "percent" as a clearly defined *number* — rather than as some ineffable concept vaguely related to "out of 100" — cannot be overemphasized. Armed with this definition, they can now think of all problems about percent as just another problem about numbers. This is an improvement over *not being able to think about any problem related to percent at all because they don't know what a "percent" is.* Naturally, the fact that school mathematics is not fond of clearcut definitions plays a role in the absence of a definition for the concept of "percent", but in this case, there is something else that works against the formulation of a definition. Observe that our definition depends on the availability of a precise concept of a *complex fraction* and, by implication, the rules in §6 that govern the computations with complex fractions. So long as school mathematics continues to pretend that complex fractions do not exist, no definition of "percent" would be possible.

By tradition, a percent $\frac{N}{100}$, where N is a fraction, is often written as $N\%$. By regarding $\frac{N}{100}$ as an ordinary fraction, we see that the usual statement $N\%$ **of a quantity** $\frac{m}{n}$ is exactly $N\% \times \frac{m}{n}$ (see the discussions at the end of §2 and at the beginning of §6). Now

$$N\% \times \frac{m}{n} \;=\; N \times \left(\frac{1}{100} \times \frac{m}{n}\right),$$

and the expression $\frac{1}{100} \times \frac{m}{n}$ means the length of 1 part when $[0, \frac{m}{n}]$ is divided into 100 equal parts. Therefore, we have proved the following lemma.

Lemma *Let N be a fraction. Then $N\%$ of $\frac{m}{n}$ is equal to N of the number A, where A is the length of 1 part when $[0, \frac{m}{n}]$ is divided into 100 equal parts.*

When N is a whole number, "N copies of A" would just have the usually meaning of N copies of a part when $[0, \frac{m}{n}]$ is divided into 100 equal parts. This is the naive concept of "percent" that most students are taught. But when N is a fraction (e.g., $\frac{3}{17}$), students are usually not told what "$N\%$ of something" means. More sobering is the fact that students are usually taught the concept of $N\%$ where N is a whole number ("N out of a hundred"), but are then asked to do computations whose answers are of the form $\frac{k}{\ell}\%$.

It is therefore worth pointing out that the interpretation in the Lemma for any fraction N *is derived strictly from the mathematics we have done.* You have not been "told" that this is true. Rather, you see for yourself that if you have mastered all the skills and definitions up to this point, then you will be led to this conclusion by mathematical reasoning.

Now, the following are examples of the three kinds of standard questions on percents that students traditionally consider to be difficult:

(i) What is 5% of 24?

(ii) 5% of what number is 16?

(iii) What percent of 24 is equal to 9?

The answers are simple consequences of what we have done *provided we follow the precise definitions.*[25] Thus, (i) 5% of 24 is $5\% \times 24 = \frac{5}{100} \times 24 = \frac{6}{5}$. For (ii), let us say that 5% of a certain number y is 16, then again strictly from the definition given above, this translates into $(5\%)y = 16$, i.e., $y \times \frac{5}{100} = 16$. By the definition of division, this says

$$y = \frac{16}{\frac{5}{100}} = 16 \times \frac{100}{5} = 320$$

[25] Always remind your students that if they don't know definitions, they are not in a position to do mathematics, in the same way that anyone who has no vocabulary is not in a position to write novels.

Finally, (iii). Suppose $N\%$ of 24 is 9. This translates into $N\% \times 24 = 9$, or $\frac{N}{100} \times 24 = 9$. Multiplying both sides by $\frac{100}{24}$, we have

$$N = \frac{900}{24} = \frac{75}{2} = 37\frac{1}{2}$$

So the answer to (iii) is $37\frac{1}{2}\%$.

What we can conclude from this short discussion is that, if students have an adequate background in fractions and are at ease with the use of symbols,[26] the concept of percent is straightforward and involves no subtlety. If this kind of instruction has been implemented in the school classroom, then education research would be in a position to shed light on what the real learning difficulties are. Until then, we should concentrate on meeting the minimum requirement of mathematics, which is to provide clear and precise definitions of all the concepts. Note however that such a definition of percent cannot be given if the concept of a complex fraction is not available.

Let us return to the problem at the beginning. There are 40 squares in the rectangle, and we must express 6 out of 40 as a percent, i.e.,

$$\text{if} \quad \frac{6}{40} = \frac{N}{100}, \quad \text{for some fraction } N\text{, what is } N?$$

It is simple:

$$N = \frac{6 \times 100}{40} = 15$$

So the answer is 15%. The teacher, however, had in mind something like this: There are 40 squares, so 4 squares constitute 10%. Another 2 would therefore add 5%. As $6 = 4 + 2$, 6 squares make up 15%.

Now the teacher's solution is not superior to the computational solution we presented above, because while it is cute, it has very limited scope. For example, how would this solution help you handle the problem of expressing 6 out of 39 squares as a percent? The computational solution does that and more. Moreover, the computation is in no way lacking in "conceptual understanding". Think for a second, and you would agree that it takes more than a little understanding to learn a precise definition of "percent", to realize the importance of the formulas (a)–(e) of §6, and to be able to prove them and use them correctly.

[26] In case you haven't noticed, we have freely made use of symbols from the beginning.

There are obvious reasons why you should learn the *precise* definition of percent, and how to use it to get the answers to all the standard questions related to this concept.

Ratio

Next we take up the concept of ratio, and it is unfortunately one that is encrusted in excessive verbiage. It would be expedient, therefore, to begin with a short definition.

Definition *Given two fractions A and B. The* **ratio of** *A* **to** *B, sometimes denoted by* $\boldsymbol{A : B}$, *is the complex fraction* $\frac{\boldsymbol{A}}{\boldsymbol{B}}$.

In connection with ratio, there are some common expressions that need to be made explicit. To say that **the ratio of boys to girls** in a classroom is 3 to 2 is to say that if B (resp., G) is *the number of* boys (resp., girls) in the classroom, then the ratio of B to G is $\frac{3}{2}$. Similarly, in making a fruit punch, the statement that **the ratio of fruit juice to rum is 7 to 2** means that we are comparing the *volumes* of the two fluids (when the use of volume as the unit is understood in this situation), and if the amount of fruit juice is A fluid ounces and the amount of rum is B fluid ounces, then *the ratio of A to B is $\frac{7}{2}$*. And so on.

We will now work out some standard problems on ratios *strictly using this definition.* The clarity of the ensuing discussion, as well as the ease with which we dispatch the problems will serve as a persuasive argument for the definition.

Example 1 In a school auditorium with 696 students, the ratio of boys to girls is 11 to 13. How many are boys and how many are girls?

Let the number of boys be B and the number of girls be G, then we are given that $\frac{B}{G} = \frac{11}{13}$. Thus by the cross-multiplication algorithm, $13B = 11G$. Let k be this common number, i.e., $13B = 11G = k$, so $B = \frac{k}{13}$ and $G = \frac{k}{11}$. Now we are also given $B + G = 696$, so $\frac{k}{13} + \frac{k}{11} = 696$. This gives $\frac{24k}{143} = 696$, and therefore $24k = 143 \times 696$, i.e., $k = 29 \times 143$. Since $B = \frac{k}{13}$, we get $B = 319$. The value of G can be obtained from either $B+G = 696$, or from $G = \frac{k}{11}$. In any case, $G = 377$.

The preceding solution is one that will be applicable in all situations. In a middle school, it would be prudent to also know a more intuitive argument that may serve as a gentler introduction. From $\frac{B}{G} = \frac{11}{13}$, we draw instead the conclusion that

$$B \ = \ \frac{11}{13} \times G$$

By the definition of multiplication (§4), this says the number B is the totality of 11 groups when the G girls are divided into 13 equal groups. Therefore the 696 students are now divided into $11 + 13$ equal groups, of which the girls comprise 13 of these groups and the boys 11. Since the size of one group is $\frac{696}{24} = 29$, we see that $G = 13 \times 29 = 377$ and $B = 11 \times 29 = 319$.

A more sophisticated problem is the following.[27]

Example 2 Divide 88 into two parts so that their ratio is $\frac{2}{3}$ to $\frac{4}{5}$.

Let the two parts be A and B. Then we are given that

$$\frac{A}{B} = \frac{\frac{2}{3}}{\frac{4}{5}}$$

Using invert-and-multiply on the right and simplifying, we get

$$\frac{A}{B} = \frac{5}{6}$$

By the cross-multiplication algorithm, $6A = 5B$. Let s be the common value. Thus $6A = s$ and $5B = s$, leading to $A = \frac{s}{6}$ and $B = \frac{s}{5}$. Because $A + B = 88$, we have $\frac{s}{6} + \frac{s}{5} = 88$, so that $\frac{11s}{30} = 88$, and therefore $s = 240$. It follows from $6A = s$ that $A = 40$, and from $5B = s$ that $B = 48$. Thus the two parts are 40 and 48.

There are other ways to solve problems such as the two preceding examples. We leave to an exercise the exploration of other methods of solution.

[27] This was a problem in a 1875 California Exam for Teachers, and it was mentioned in the well-known address of Lee Shulman, "Those who understand: Knowledge growth in teaching," *Educational Researcher* 15 (1986), 4-14.

The concept of constant rate

In school mathematics, the most substantial application of the concept of division is to problems related to *rate*, or more precisely, *constant rate.* The precise definition of the general concept of "rate" requires more advanced mathematics, and in any case, it is irrelevant whether we know what a *rate* is or not.[28] What is relevant is to know the precise meaning of "constant rate" in specific situations, and the most common of these situations will be enumerated in due course. Among these, the "rate" involving motion is what we call *speed.* Because this concept may be the most intuitive, we proceed to discuss it in some detail.

Instead of giving, outright, a definition of what constant speed motion is, we begin with the more basic concept of the **average speed over a time interval from time $\boldsymbol{t_1}$ to time $\boldsymbol{t_2}$**, $t_1 < t_2$, as

$$\frac{\text{distance traveled from } t_1 \text{ to } t_2}{t_2 - t_1}$$

> *What needs to be singled out is the fact that the term "average speed" by itself carries* no *information, because we have to know the average speed* from a specific point in time t_1 to another point in time t_2. *In addition, because the terminology ("average") stimulates the conditioned reflex of "add two numbers and divide by 2", students need to put this conditioned reflex in check. The added cognitive complexity associated with "average speed" is thus something your students will not take to kindly when you teach this concept, but it is nevertheless something you must impress on them because mastering subtleties of this kind prepares them for higher mathematics.*

We want to show that if a motion has a fixed average speed regardless of the time interval, then it is, intuitively, what we call "constant speed". For example, the intuitive meaning of constant speed is that over time intervals of the same length,

[28] Students of calculus beware! If you want to be able to teach school mathematics, you still need to learn how to explain *constant rate* in an elementary manner without once mentioning "a function whose derivative is a constant". This is in fact an excellent example of why, no matter how much advanced mathematics one knows, one must learn *school mathematics* in order to be a good teacher.

the distances traveled are the same. Now if we have a motion with a fixed average speed v, we are going to prove that it has the same property. Let us say the two time intervals $[t_1, t_2]$ and $[t_3, t_4]$ have the same length so that $t_2 - t_1 = t_4 - t_3$. Then if d and d' are the distances traveled during the time intervals $[t_1, t_2]$ and $[t_3, t_4]$, respectively, by this motion with fixed average speed v, we have

$$v = \frac{d}{t_2 - t_1} = \frac{d'}{t_4 - t_3}$$

so that

$$\begin{aligned} d &= v(t_2 - t_1) \\ &= v(t_4 - t_3) \qquad\qquad \text{(because } t_2 - t_1 = t_4 - t_3\text{)} \\ &= d' \end{aligned}$$

as desired. Next, for a motion of constant speed, if the time interval $[s_1, s_2]$ is $\frac{1}{n}$ as long as the time interval $[t_1, t_2]$, then the distance traveled in $[s_1, s_2]$, say D, is also expected to be $\frac{1}{n}$ of the distance traveled in $[t_1, t_2]$, say d. We now show that this is also true of motions with a fixed average speed v. This is because if we divide $[t_1, t_2]$ into n equal parts, then the length of each of these parts is just $s_2 - s_1$, by assumption, so that by what we have just proved, the distance traveled during $[s_1, s_2]$ is the same as the distance traveled during each of these parts, i.e., D. Since the total distance traveled in all n parts is exactly the distance traveled in $[t_1, t_2]$, which is d, we have:

$$\underbrace{D + D + \cdots + D}_{n} = d.$$

It follows that $D = \frac{1}{n}\,d$, as claimed.

The following definition is now seen to be entirely reasonable.

Definition *A motion is said to have* **constant speed** *v if the average speed of the motion over* any *time interval is v.*

Why do we need a *definition* of constant speed? After all, is it not enough to know that, during any one-hour interval (or any one-minute interval), the distance traveled is always the same? The answer if no. First of all, the perception that no definition of constant speed is needed is a reflection of the fact that all constant speed or constant rate problems are usually done in school classrooms "by common sense" and without any logical deductions. But what is "common sense" to one person may not be "common sense" to another, especially if the other person is a beginner unsure of himself or herself. A solution that depends on some ineffable, and not necessarily universal, understanding of the situation is not learnable mathematics. Let us illustrate. The following is a standard problem.

> If Ina can walk $3\frac{2}{5}$ miles in 90 minutes, how long would it take her to walk half a mile?

A common solution would be like this: Suppose it takes Ina x minutes to walk half a mile, then proportional reasoning shows that $3\frac{2}{5}$ is to $\frac{1}{2}$ as 90 is to x. So

$$\frac{3\frac{2}{5}}{\frac{1}{2}} = \frac{90}{x}$$

By the cross-multiplication algorithm for *complex* fractions ((b) of §6),[29]

$$x = \frac{1}{3\frac{2}{5}} \times \left(\frac{1}{2} \times 90\right) = \frac{45}{\frac{17}{5}} = 13\frac{4}{17} \text{ minutes}$$

The answer is undoubtedly correct, but how is anyone supposed to explain the "proportional reasoning" to an eleven-year-old so that she can learns how to form the proportion *with conviction?* And constant speed is not even assumed! This solution is mathematically senseless. Let us do it over again. We begin with a reformulation of the problem that at least makes sense.

> If Ina walks at a constant speed and she walks $3\frac{2}{5}$ miles in 90 minutes, how long would it take her to walk half a mile?

Solution Let Ina walk at the constant speed of v mph, and assume that it takes her x minutes to walk half a mile. Then her average speed over the time interval of 90 minutes and that over the time interval of x minutes are both v mph, by assumption of constant speed. Thus

$$\frac{3\frac{2}{5}}{90} = v = \frac{\frac{1}{2}}{x} \qquad (6)$$

[29] Take note of the very natural way a formula about complex fractions in §6 is called upon to serve a mathematical need.

So the the cross-multiplication algorithm for complex fractions ((b) of §6) gives

$$x \;=\; \frac{1}{3\frac{2}{5}} \times \left(\frac{1}{2} \times 90\right) \;=\; 13\frac{4}{17} \text{ minutes}$$

It is the same answer, but now every step is logical, and what is more important, we finally see *how the assumption of constant speed becomes an integral part of the solution.* At least, if someone wants to learn it, she can learn it step-by-step, without appealing to any "supernatural common sense".

Let us now finish the answer to the question of whether, instead of having a precise definition of constant speed, it suffices to know that during any one-hour interval (or any one-minute interval), the distance traveled is always the same. Look at the preceding solution: the critical step in equation (6) requires that we know that Ina's average speed over the time interval x is still v, *and we have no idea what x is.* So, no, it is not enough to know that for a motion of constant speed, the speed over a fixed time interval (regardless of what it is, an hour, a minute, or a second) is a fixed number. We need the full definition of *constant speed.*

In the language of school mathematics, speed is the "rate" at which the work of moving from one place to another is done. There are other standard "rate" problems which deserve to be mentioned. One of them is painting (the exterior of) a house. The rate there would be the number of square feet painted per day or per hour. A second one is mowing a lawn. The rate in question would be the number of square feet mowed per hour or per minute. A third is the work done by water flowing out of a faucet, and the rate is the number of gallons of water coming out per minute or per second. In each case, the concept of **constant rate** can be precisely defined as in the case of constant speed. For example, the concept of **constant rate of lawn-mowing** can be defined by first defining the **average rate of lawn-mowing** from time T_1 to time T_2 as $\frac{A'}{T_2 - T_1}$, where A' is the area mowed from time T_1 to time T_2. Then the lawn is said to be **mowed at a constant rate** if the average rate of the lawn being mowed over *any* time interval $[T_1, T_2]$ is equal to a fixed constant.

However, textbooks over the years have developed an "abstract" kind of work problems, which typically read as follows.

> It takes Regina 10 hours to do a job, and Eric 12 hours. If they work together, how long would it take them to get the job done?

The mathematical defects of such a problem are overwhelming. First, this problem cannot be solved if Regina and Eric do not each work at a constant rate, yet *the assumption of constant rate is typically not mentioned.* A second assumption is that, somehow, Regina and Eric manage to do different parts of the job, and at the end the two parts fit together perfectly to get the job done faster. If the nature of the work is not made explicit, however, such an assumption would sorely tax a student's imagination. For example, suppose the job involved is driving from Town A to Town B, and a student interprets "working together" to mean Regina and Eric sharing the driving! A third serious defect is that the concept of constant rate becomes difficult to formulate *precisely* when the job in question is not clearly specified. Indeed, the average rate of work from time t_1 to time t_2 is by definition,

$$\frac{\text{the amount of work done from } t_1 \text{ to } t_2}{t_2 - t_1}$$

But the numerator has to be a *number*, and a student would have a hard time associating the vague description of "amount of work" with a number. Such vagueness interferes with the learning of mathematics.

> **Make sure that you do not damage your students' learning by teaching them only this kind of generic "work problems".**

It is conceivable that after doing many well formulated work problems, students are already so familiar with the general reasoning that they can afford to take shortcuts by engaging in doing such abstract work problems. This is something you as a teacher has to decide, but certainly, this kind of problems are inappropriate for beginners.

Units and dimension analysis

Many teachers are concerned about getting students to use the correct unit in such rate problems. As a result of this concern, something called "dimension analysis" has sprung up to help students learn about changing one unit to another. Dimension analysis *is* used extensively in science and engineering as a quick check on the correct use of units because one can imagine that in physics, for example, all kinds of units have to be used to fit the occasion. Thus for the study of a motion within the lab in a time interval of 3 seconds, one might have second thoughts about using *miles per hour*; perhaps *feet per second* or *meters per second* would be more appropriate. But even in physics, dimension analysis cannot replace the knowledge of why a unit of acceleration is m/sec^2 or a unit of momentum is kg-m/sec. There is need for a basic understanding of the processes involved. Why this is relevant is that, in school mathematics, dimension analysis is taught as a rote skill. While it is possible to explain the procedures used in dimension analysis, any minute spent on such an explanation is a minute taken away from acquiring an understanding the underlying mathematics of constant rate. Furthermore, if students know the definitions and can follow the definitions faithfully, then they will see that there is no mystery to changing units. In this short sub-section, we illustrate this point of view with two examples.

One suggestion for minimizing the distraction of dimension analysis is to agree, from the beginning, on a single unit of measurement. For example, if a problem involves motion and the data are given in minutes and hours in time, and miles, yards, and meters in length, then it would make sense to convert everything into hours and miles only. In the subsequent computations, there would be no need to worry about unit conversion.

Suppose water comes out of a faucet at a constant rate of 5 gallons per minute. We show how to express this rate in terms of quarts and seconds. In other words, how many quarts of water come out of the faucet each second? Recall that the meaning of *constant* rate is that the average rate over any time interval is the same number, so we look at the average rate of the water flow in a time interval of one minute. We use a one-minute interval because are given that in this time interval, 5 gallons come out of the faucet. Now one minute is 60 seconds, and each gallon is 4 quarts. Since 5 gallons is $5 \times 4 = 20$ quarts, we are therefore given that 20 quarts come out of the faucet every 60 seconds. The *average rate* in a 60 second time interval is, *by definition*, the quotient

$$\frac{20 \text{ (quarts)}}{60 \text{ (seconds)}} = \frac{20}{60} \text{ quarts per second } = \frac{1}{3} \text{ quarts/second}$$

Since we are assuming constant rate, we see that this average rate is in fact the constant rate, i.e., $\frac{1}{3}$ qt./sec.

Once we are more used to this reasoning, we would do the conversion directly without further ado, as follows:

$$5 \text{ gal./min.} \;=\; \frac{5 \times 4 \text{ qt.}}{60 \text{ sec.}} \;=\; \frac{20}{60} \text{ qt./sec.} \;=\; \frac{1}{3} \text{ qt./sec.}$$

As another example, suppose an object travels at a constant speed of $85\frac{1}{5}$ ft./sec. What is its speed in terms of mph (miles per hour)? We know 1 mile = 5280 ft., so $85\frac{1}{15}$ ft. $= 85\frac{1}{15} \times \frac{1}{5280}$ mi.$= \frac{29}{1800}$ mi. On the other hand, 1 hour is 3600 seconds, so that 1 second is $\frac{1}{3600}$ hour. The object therefore travels $\frac{29}{1800}$ mi. in a time interval of $\frac{1}{3600}$ hr. By definition of average rate, the average rate of the motion in a time interval of $\frac{1}{3600}$ hr. is the quotient

$$\frac{\frac{29}{1800}}{\frac{1}{3600}} \text{ mph} \;=\; 58 \text{ mph}$$

Again, having gone through this process once, we can now compute more simply:

$$85\frac{1}{15} \text{ ft./sec.} \;=\; \frac{85\frac{1}{15} \times \frac{1}{5280}}{1 \times \frac{1}{3600}} \text{ mph} \;=\; 58 \text{ mph}$$

Cooperative work

We now revisit the previous example about Reginal and Eric by giving four different reformulations that are all mathematically acceptable:

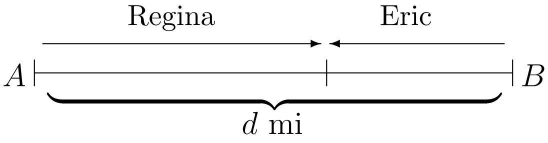
(P1) Regina drives from Town A to Town B in 10 hours, and Eric in 12. Assuming that each drives at the same constant speed, Regina from Town A to Town B, and Eric from Town B to Town A, and that they drive on the same highway, after how many hours will they meet in between?

(P2) Regina mows a lawn in 10 hours, and Eric in 12. Assuming that each mows at the same constant rate, how long would it take them to mow the same lawn if they mow together without interfering with each other?

(P3) Regina paints a house in 10 hours and Eric in 12. Assuming that each paints at the same constant rate, how long would it take them to paint the same house if they paint together without interfering with each other?

(P4) A faucet can fill a tub in 10 minutes, and a second faucet in 12. Assuming that the rate of the water flow remains constant in each faucet, how long would it take to fill the same tub if both faucets are turned on at the same time?

It should be recognized that *all four problems are the same problem*: if you can solve one, you can solve them all. Let us give a solution of the first, (P1).



We have to determine the speeds of Regina and Eric. We do not know the distance between Towns A and B, so to facilitate thinking, let us say this distance is d miles. Therefore Regina's speed v_R satisfies $d = 10v_R$, and we have $v_R = \frac{d}{10}$ mph. Similarly, Eric's speed v_E is $\frac{d}{12}$ mph. We have to find out how long it takes Regina and Eric to meet, but again, to facilitate thinking, let us say Regina and Eric meet after T hours. At the moment we do not know what T is, but the assumption of constant speed guarantees that the distance Regina has driven in T hours is $v_R T = \frac{dT}{10}$ miles. Similarly, the distance Eric has driven after T hours is $\frac{dT}{12}$ miles. Since they meet in between the towns, the total distance they have driven together after T hours is exactly d miles. Therefore we have

$$\frac{dT}{10} + \frac{dT}{12} = d$$

By the distributive law (e) for complex fractions in §6, we have

$$dT\left(\frac{1}{10} + \frac{1}{12}\right) = d$$

Since d is just a number, multiplying both sides by the complex fraction $\frac{1}{d}$ (and using rules (a) and (d) of §6) gives $\left(\frac{1}{10} + \frac{1}{12}\right)T = 1$. By the definition of division. we get

$$T = \frac{1}{\frac{1}{10} + \frac{1}{12}} = 5\frac{5}{11} \text{ (hours)}$$

It may be instructive if we also solve problem (P2) for comparison.

Let the area of the lawn be A sq. ft. Because in 10 hours Regina can mow the whole lawn, i.e., A sq. ft., her (constant) rate of lawn-mowing is, by definition, $\frac{A}{10}$ sq. ft. per hour. Similarly, Eric's rate of lawn-mowing is $\frac{A}{12}$ sq. ft. per hour. Now suppose the two together can finish mowing the lawn in T hours. If in T hours, Regina mows R sq. ft., then by definition of constant rate, $\frac{R}{T} = \frac{A}{10}$, and therefore, $R = \frac{AT}{10}$. Similarly, in T hours, Eric mows $\frac{AT}{12}$ sq. ft. Because they mow with no interference from each other, the sum total of the areas they mow in T hours adds up exactly to A, i.e.,

$$\frac{AT}{10} + \frac{AT}{12} = A$$

By the distributive law, $AT(\frac{1}{10} + \frac{1}{12}) = A$. Multiplying both sides by $\frac{1}{A}$, we get $T(\frac{1}{10} + \frac{1}{12})= 1$, so that

$$T = \frac{1}{\frac{1}{10} + \frac{1}{12}} = 5\frac{5}{11} \text{ (hours)},$$

exactly as before.

The next example is slightly more intricate.

Example 3 Tom and May drive on the same highway at constant speed. May starts 30 minutes before Tom, and her speed is 45 mph. Tom's speed is 50 mph. How many hours *after May leaves* will Tom catch up with her?

We give two slightly different solutions. Suppose T hours after May leaves, Tom catches up with May. In those hours, May has driven $45T$ miles. Since Tom does not start driving until half an hour after May does, the total distance he travels in that time duration is $50(T - \frac{1}{2})$ miles. The two distances being equal, we get $45T = 50(T - \frac{1}{2})$. By the distributive law, $45T = 50T - 25$. Adding 25 to both sides, we get $45T + 25 = 50T$, and so we get $25 = 5T$ after subtracting $45T$ from both sides. Thus $T = 5$, i.e., 5 hours after May leaves, Tom catches up with her.

Another solution is to watch Tom's car from May's car. So starting at $\frac{1}{2}$ hour after she leaves, she sees Tom's car coming from a distance of $45 \times \frac{1}{2} = 22.5$ miles.[30] Let t measure the number of hours after Tom starts driving. In t hours, Regina's car

[30] She has omnidirectional vision.

travels $45t$ miles, whereas Tom's car travels $50t$ miles. Therefore, after t hours, May's observation is that Tom is $50t - 45t = 5t$ miles closer to her car, which is the same as saying that, Tom's car — as observed from May's car — travels $5t$ miles in t hours. By definition of constant speed, Tom's car is driving at a constant speed of 5 mph when observed from May's car. Since May is initially 22.5 miles away, it will take Tom $\frac{22.5}{5} = 4.5$ hours to catch up. Since Tom starts 0.5 hours after May leaves, it takes Tom $4.5 + 0.5 = 5$ hours after May leaves to catch up with her.

As a final example, we give one that is the kind of brain-teaser that one passes around in parties. The emphasis will not be in the answer but in the method of solution: If you are explaining to seventh graders how to solve this, you will not be trying to "make an impression" in a party but will try instead to teach mathematics. How can you explain it *logically* when the obvious, illegal method of solution is so seductive and so "simple"?

EXAMPLE 4 If 3 people can paint 4 houses in 5 hours, and if everybody paints at the same constant rate and all the houses are identical, how long will it take 2 people to paint 5 houses? (Assume that they don't interfere with each other's work.)

First, the illegal solution. One person can paint $\frac{4}{3}$ houses in 5 hours, and therefore can paint 1 house in

$$\frac{5}{\frac{4}{3}} = \frac{15}{4} \text{ hours}$$

So 2 people can paint 1 house in $\frac{1}{2} \times \frac{15}{4} = \frac{15}{8}$ hours. Therefore, to paint 5 houses, it takes 2 people $2 \times \frac{15}{8} = \frac{75}{8}$ hours. *Again, the answer is correct, but this method of solution has a fatal flaw: the assumption of "constant rate of painting" is never explicitly used. Since clearly this assumption is crucial for the solution, your job as a teacher is to de-construct this glib solution to bring out the constancy assumption. Only then can you hope that your students will learn from it.*

We try again. Let each house have A sq. ft. of surface area for painting, and let everyone paint at the constant rate of r sq. ft. /hr. We claim that the number of sq. ft. painted by k people (where k is a positive integer) in t hours is krt sq. ft.

Indeed, by definition,

$$\frac{\text{\# sq. ft. painted by one person in } t \text{ hrs.}}{t \text{ hrs.}} \;=\; r \text{ sq. ft./hr.},$$

each person paints rt sq. ft. in t hours. But they all paint at the same rate r, so k people in t hours will paint $(rt+rt+\cdots+rt)$ (k times) sq. ft., which is krt sq. ft., as claimed. It follows that the rate at which k people (with k fixed during the discussion) paint together is *a constant*, namely, kr sq. ft./hr. This is because, by definition, the *average* rate at which k people paint in a given time interval of t hours is

$$\frac{\text{\# sq. ft. painted by } k \text{ people in } t \text{ hours}}{t \text{ hours}} \;=\; \frac{krt}{t} \;=\; kr \text{ sq. ft./hr.} \qquad (7)$$

This average rate being constant (independent of t), this is exactly what we mean by *constant rate*. Now we know that 3 people can paint 4 houses in 5 hours. Therefore equation (7) gives

$$\frac{\text{\#sq. ft. painted by 3 people in 5 hours}}{\text{5 hours}} \;=\; 3r \text{ sq. ft./hr}$$

But 4 houses have $4A$ sq. ft. of total area, so

$$\frac{4A}{5} \;=\; 3r,$$

and we obtain an explicit value of r:

$$r \;=\; \frac{4A}{15} \text{ sq. ft./hr.}$$

Let us say it takes T hours for 2 people to paint 5 houses, then by equation (7) again,

$$\frac{\text{\# sq. ft. painted by 2 people in T hours}}{T \text{ hours}} \;=\; 2r \text{ sq. ft./hr}$$

which is to say,

$$\frac{5A}{T} \;=\; 2 \times \frac{4A}{15},$$

and we finally get

$$T \;=\; \frac{5}{\frac{8}{15}} \;=\; \frac{75}{8} \text{ hours}$$

Exercises 1.7

A special word of caution for this set of problems. Most of these problems are likely known to you, and therefore you have inherited their template solutions from your school days or from the textbooks you use. Equally likely, these template solutions don't make mathematical sense. You should therefore approach these problems with fresh eyes and make a new beginning. Write out their solutions the way you are going to teach your students, and then ask yourself: Do I make sense to my students? Make sure you can answer this question in the affirmative.

1. A hi-fi store sells a CD player for \$225. The owner decides to increase sales by not charging customers the 8% sales tax. Then he changes his mind and charges customers $\$x$ so that, after they pay the sales tax, the total amount they pay is still \$225. What is x?

2. Helena drives from Town A to Town B at x mph, and drives back at y mph. What is her *average speed* for the round trip? If the round trip takes t hours, how far apart are the towns?

3. A high-tech stock dropped 45% of its value in June to its present value of \$N. A stock broker tells his clients that if the stock were to go up by 60% of its present value, then it would be back to where it was in June. Is he correct? If so, why? If not, by what percent must the stock at its present value of \$N rise in order to regain its former value?

4. A fully open faucet (with a constant rate of water flow) takes 25 seconds to fill a container of $5\frac{1}{2}$ cubic feet. At the same rate, how long does it take to fill a tank of $12\frac{1}{2}$ cubic feet? (*Be careful with your explanation!*)

5. A faucet with a constant rate of water flow fills a tub in 9 minutes. If the rate of water flow were to increase by 10%, how long would it take to fill the tub?

6. Kate and Laura walk straight toward each other at constant speed. Kate walks $1\frac{2}{3}$ times as fast as Laura. If they are 2000 feet apart initially, and if they meet after $2\frac{1}{2}$ minutes, how fast does each walk?

7. (Sixth-grade Japanese exam question) A train 132 meters long travels at 87 kilometers per hour and another train 118 meters long travels at 93 kilometers per hour. Both trains are traveling in the same direction on parallel tracks. How many seconds does it take from the time the front of the locomotive of the faster train reaches the end of the slower train to the time that the end of the faster train reaches the front of the locomotive on the slower one?

8. Solve Example 2 of the sub-section on ratio by a method similar to the second solution of Example 1. Can you find other methods of solution to both Examples 1 and 2?

9. Driving at her usual constant speed of v mph, Stefanie can get from point A to point B in 5 hours. Today, after driving 1 hour, she decides to speed up to a constant speed of w mph so that she can finish the whole trip in $4\frac{1}{2}$ hours instead of 5. By what percent is w bigger than v (compared with v)?

10. Winnie and Reggie working together can paint a house in 56 hours. If Reggie paints the same house alone, it takes him 90 hours to get it done. How long does it take Winnie to paint the house if she works alone? (Assume each paints at a constant rate, and that when they paint together there is no mutual interference.)

11. Alfred, Bruce, and Chuck mow lawns at a constant rate. It takes them 2 hours, 1.5 hours, and 2.5 hours, respectively, to finish mowing a certain lawn. If they mow the same lawn at the same time, and if there is no interference in their work, how long will it take them to get it done?

12. How much money would be in an account at the end of two years if the initial deposit was \$93 and the bank pays an interest of 6% at the end of each year? (This means, if there are n dollars in the account at the end of the year, then the bank adds $6\% \times n$ dollars into the account.) And at the end of n years?

13. If A, B, C are three numbers which satisfy $A : B = \ell : m$ and $B : C = m : n$ for some fractions ℓ, m, n, then we abbreviate by writing $\boldsymbol{A : B : C = \ell : m : n}$ to

express the equality of the two pairs of ratios, and say that A, B, C, are **in the ratio $\boldsymbol{\ell : m : n}$**. Prove that A, B, C, are in the ratio $\ell : m : n$ if and only if $\frac{A}{\ell} = \frac{B}{m} = \frac{C}{n}$.

14. Benoit, Carl, and Davida chip in to buy a hi-fi system. The cost is \$434, and their contributions (in the order of Benoit, Carl, and Davida) are in the ratio of $2 : 7 : 4$ (see problem 13). How much does each contribute?

15. (a) If 5 people can paint 7 houses in 3 days, how long would it take 2 people to paint 9 houses? (b) If 4 people can paint 3 houses in 2 days, how many houses can 5 people paint in 6 days? (In both cases, assume that everyone paints at a constant rate, they never interfere with each other, and all the houses are identical.)

16. On June 16, 2009, Reuters made the following report:

> MySpace, the social network owned by Rupert Murdoch's News Corp, said it will cut 30 percent of its staff to lower costs as it struggles to stay popular in the face of rising competition.
>
> MySpace will be left with about 1,000 employees, it said in a statement released on Tuesday. The company declined to say how many people work at the service, but the percentage suggests that about 400 people will lose their jobs.

(a) Compute roughly how many people were working at MySpace at the time of the report. (b) Would you care to make any social commentary on this report? (Subsequently, it was revealed that MySpace had 1420 employees at the time.)

Appendix

In this appendix, we briefly recall the commutative and associative laws of addition and multiplication, and also the distributive law that connects the two. Some standard consequences will also be discussed.

In the following, lowercase italic letters will be used to stand for arbitrary numbers without further comment. Notice that we are intentionally vague about what "numbers" we are talking about. The fact is that Theorems 1 and 2 are valid for whole numbers, integers, rational numbers, real numbers, and even complex numbers, and these theorems will be used in such generality without comment for the rest of these notes. With this understood, the **associative** and **commutative laws for addition** state that for any x, y, z, we always have

$$x + (y + z) = (x + y) + z$$

and

$$x + y = y + x,$$

respectively. A fairly tedious argument, one that is independent of the specific numbers x, y, z involved but is dependent *formally* only on these two laws, then leads to the following general theorem. For everyday applications, this theorem is all that matters as far as these simple laws are concerned:

Theorem 1 *For any finite collection of numbers, the sums obtained by adding them up in any order are all equal.*

A similar discussion holds for multiplication. Thus the **associative** and **commutative laws for multiplication** state that for any x, y, z, we always have

$$x(yz) = (xy)z$$

and

$$xy = yx,$$

respectively. And, in like manner, we have:

Theorem 2 *For any finite collection of numbers, the products obtained by multiplying them in any order are all equal.*

Finally, the **distributive law** is the link between addition and multiplication. It states that, for any x, y, z,

$$x(y + z) = xy + xz$$

Here it is understood that the multiplications xy and xz are performed before the products are added. A simple argument then extends this law to allow for any number of additions other than two. For example, the distributive law for five additions states that for any x, a, b, c, d, e, we have

$$x(a+b+c+d+e) = xa+xb+xc+xd+xe$$

Chapter 2: Rational Numbers

We are going to revisit the number line. Up to now, we have only made use of the right side of 0. It is time that we make full use of the entire number line, both to the left and right. Because we already have the fractions to the right of 0, we now look at the collection of numbers (i.e., points on the number line) to the left of 0 obtained by reflecting the fractions across 0. The fractions together with their reflected images will be seen to form a number system, in the sense that we can perform the four arithmetic operations on them in a way that is consistent with the operations already defined on the fractions. This number system, called the *rational numbers*, is the subject of this chapter.

1 The two-sided number line

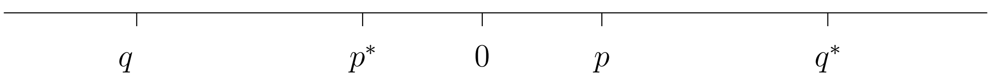
Recall that a **number** is a point on the number line. We now look at all the numbers as a whole. Take any point p on the number line which is not equal to 0; such a p could be on either side of 0 and, in particular, does not have to be a fraction. Denote the **mirror reflection of p** on the opposite side of 0 by p^*, i.e., p and p^* are **equi-distant** from 0 (i.e., same distance from 0) and are on opposite sides of 0. If $p = 0$, let

$$0^* = 0$$

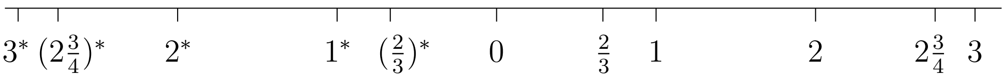
Then for any point p, it is clear that

$$p^{**} = p$$

This is nothing but a succinct way of expressing the fact that reflecting a nonzero point across 0 twice in succession brings it back to itself (if $p = 0$, of course $0^{**} = 0$). Here are the mirror reflections of two points p and q on the number line:



Because the fractions are to the right of 0, the numbers such as 1^*, 2^*, or $(\frac{9}{5})^*$ are to the left of 0. Here are some examples of the mirror reflections of fractions (remember that fractions include whole numbers):



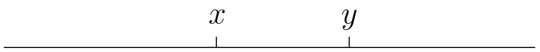
The set of all the fractions and their mirror reflections, i.e., the numbers $\frac{m}{n}$ and $(\frac{k}{\ell})^*$ for all whole numbers k, ℓ, m, n ($\ell \neq 0$, $n \neq 0$), is called the **rational numbers**, and is denoted by $\mathbf{Q}$. Recall that the whole numbers, denoted by $\mathbf{N}$, are a sub-set of the fractions. The set of whole numbers and their mirror reflections,

$$\ldots 3^*,\ 2^*,\ 1^*,\ 0,\ 1,\ 2,\ 3, \ldots$$

is called the **integers**, and is denoted by $\mathbf{Z}$. If we employ the standard symbol $\subset$ to denote **"is contained in"**, then we have:

$$\mathbf{N} \subset \mathbf{Z} \subset \mathbf{Q}$$

We now extend the **order** among numbers from fractions to all numbers: for any x, y on the number line, $\boldsymbol{x < y}$ means that x is to the left of y. An equivalent notation is $\boldsymbol{y > x}$.



Numbers which are to the right of 0 (thus those x satisfying $x > 0$) are called **positive**, and those which are to the left of 0 (thus those that satisfy $x < 0$) are **negative**. So 2^* and $(\frac{1}{3})^*$ are negative, while all nonzero fractions are positive. The mirror reflection of a positive number is therefore negative, by definition, but the mirror reflection of a negative number is positive. *The number 0 is, by definition, neither positive nor negative.*

You are undoubtedly accustomed to writing, for example, 2^* as -2 and $(\frac{1}{3})^*$ as $-\frac{1}{3}$. You also know that the "$-$" sign in front of -2 is called the **negative sign**. So you may wonder why we employ this $*$ notation and have avoided mentioning the negative sign up to this point. The reason is that the negative sign, having to do with the operation of subtraction, simply will not figure in our considerations until we begin to subtract rational numbers. Moreover, the terminology of "negative sign" carries certain psychological baggage that may interfere with learning rational numbers the proper way. For example, if $a = -3$, then there is nothing "negative" about $-a$, which is 3. It is therefore best to hold off introducing the negative sign until its natural arrival in the context of subtraction in the next section.

Exercises 2.1

1. Show that between any two rational numbers, there is another rational number.

2. Which is bigger? $(1.23)^*$ or $(1.24)^*$? $(1.7)^*$ or $\left(\frac{12}{7}\right)^*$? $(587\frac{1}{5})^*$ or $(587\frac{2}{11})^*$? $(\frac{9}{16})^*$ or $(\frac{4}{7})^*$?

3. Which of the following numbers is closest to 0 (on the number line)?

$$\left(\frac{15}{7}\right)^*, \quad \left(\frac{11}{5}\right)^*, \quad \frac{13}{6}, \quad \frac{9}{4}$$

2 Adding rational numbers

Why rational numbers?

Before we proceed to a discussion of the arithmetic operations with rational numbers, we should ask why we bother with rational numbers at all. To answer this question, we first take a backward step and look at the transition from whole numbers to fractions: with the whole numbers at our disposal, why did we bother with fractions? One reason is to consider the problem of solving equations. If we ask which *whole number* x has the property that when multiplied by 7 it equals 5, the answer is obviously "none". The fraction $\frac{5}{7}$, on the other hand, has exactly this property. One may therefore say that if we insist on getting a solution to the equation $7x = 5$, then we would inevitably be led to $x = \frac{5}{7}$. More generally, the solution to the equation $nx = m$ where m, n are given whole numbers, with $n \neq 0$, is $\frac{m}{n}$. In this sense, we may regard the fractions as the numbers which are the solutions of the equation $nx = m$ with $n \neq 0$, as m and n run through all whole numbers. Once we have these new numbers, then we introduce the arithmetic operations among them in a way that is consistent with the original arithmetic operations among whole numbers. Recall in this connection that we were careful all through Chapter 1 to stress the fact that each arithmetic operation on fractions is derived directly from the corresponding one on whole numbers.

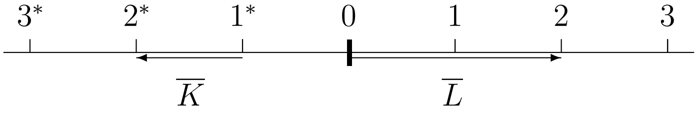
We now come back to our present situation. With fractions at our disposal, suppose we want a fraction x so that $\frac{2}{3} + x = 0$. We would get no solution in this case. The number $(\frac{2}{3})^*$ is there precisely to provide a solution to this equation. In the same way, the number $(\frac{m}{n})^*$ will be designated to be the solution of the equation $\frac{m}{n} + x = 0$, for any whole numbers m and n ($n \neq 0$). Now that we have the negative fractions, we are faced with the same problem concerning the rational numbers that we faced concerning the fractions, namely, how to define the arithmetic operations among the rational numbers in a way that is consistent with the original arithmetic operations among fractions. In this section we deal with addition.

We will approach the addition of rational numbers by imitating what we did in Chapter 1, which is to explicitly define the sum of two rational numbers and then show that, so defined, it coincides with the usual addition of fractions when the rational numbers are themselves fractions. This is not the only way to proceed, but in a middle school classroom, such a concrete approach may be best. We begin by introducing a new concept.

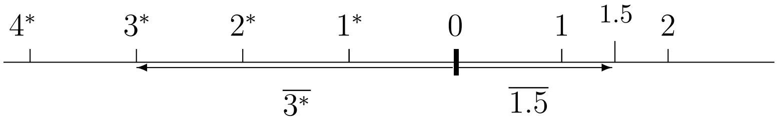
Vectors and vector addition

By definition, a **vector** is a segment on the number line together with a designation of one of its two endpoints as a **starting point** and the other as an **endpoint**. We will continue to refer to the length of the segment as the **length of the vector**, and call the vector **left-pointing** if the endpoint is to the left of the starting point, **right-pointing** if the endpoint is to the right of the starting point. The **direction** of a vector refers to whether it is left-pointing or right-pointing.

We denote vectors by placing a bar above the letter, e.g., $\overline{A}$, $\overline{x}$, etc., and in pictures we put an arrowhead at the endpoint of a vector to indicate its direction. For example, the vector $\overline{K}$ below is left-pointing and has length 1, with a starting point at 1^* and an endpoint at 2^*, while the vector $\overline{L}$ is right-pointing and has length 2, with a starting point at 0 and an endpoint at 2.



For the purpose of discussing the addition of rational numbers, we can further simplify matters by restricting attention to a special class of vectors. Let x be a number (a point on the number line), then we define the vector $\overline{x}$ to be the vector with starting point at 0 and endpoint at x. It follows from the definition that, *if x is positive, then the segment of the vector $\overline{x}$ is exactly $[0, x]$.*[31] Here are two examples of vectors arising from rational numbers: $\overline{3^*}$ and $\overline{1.5}$.

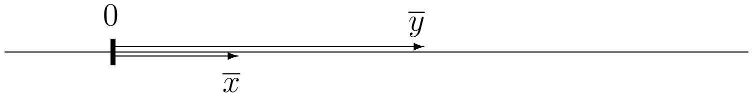


In the following, we will concentrate only on vectors of the type $\overline{\boldsymbol{x}}$ where $\boldsymbol{x}$ is a number, so that all vectors under discussion will have starting point at 0. We now describe how to add such vectors. Given $\overline{x}$ and $\overline{y}$, where x and y are two numbers, the **sum vector $\overline{\boldsymbol{x}} + \overline{\boldsymbol{y}}$** is, by definition, the vector whose starting point is 0, and whose endpoint is obtained as follows:

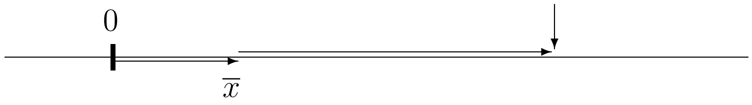
> *slide the vector $\overline{y}$ along the number line until its starting point (which is 0) is at the endpoint of $\vec{x}$, then the endpoint of $\overline{y}$ in this new position is by definition the endpoint of $\overline{x} + \overline{y}$.*

As an example, suppose we are given the following vectors $\overline{x}$ and $\overline{y}$:

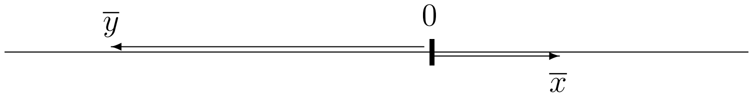
[31] In contrast with the fact that if x is negative, then the segment of $\overline{x}$ would be $[x, 0]$.



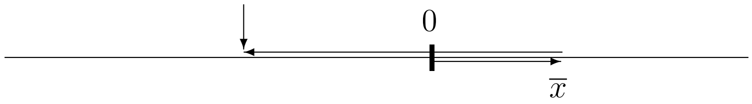
Then the vertical arrow on the right indicates the endpoint of the sum vector $\overline{x} + \overline{y}$:



On the other hand, if we are given the following two vectors $\overline{x}$ and $\overline{y}$,



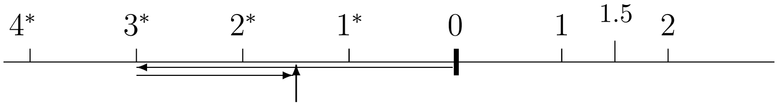
then the endpoint of $\overline{x} + \overline{y}$ is indicated by the vertical arrow on the left:



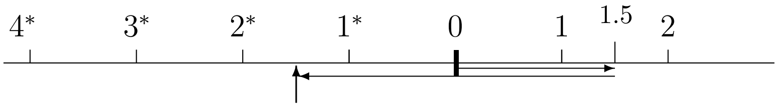
We want to show that the addition of vectors is commutative, i.e.,

$$\overline{x} + \overline{y} \;=\; \overline{y} + \overline{x} \qquad \text{for all } x, y \in \mathbf{Q}$$

Consider the sum of the previous two vectors, $\overline{3^*} + \overline{1.5}$. The endpoint of the sum vector is indicated by the vertical arrow below the number line.



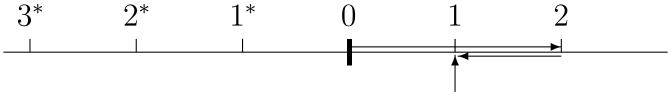
And here is the sum of the same two vectors, but with order reversed: $\overline{1.5} + \overline{3^*}$. The endpoint of this sum vector is again indicated by the vertical arrow below the number line.



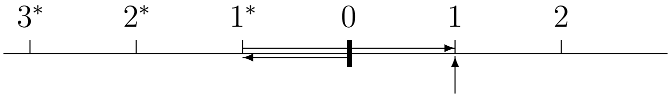
We have to show that the vectors $\overline{3^*} + \overline{1.5}$ and $\overline{1.5} + \overline{3^*}$ are equal. Because the directions of $\overline{3^*}$ and $\overline{1.5}$ are different, the definition of vector addition implies that the length of either sum vector has to be the *difference* of the lengths of $\overline{3^*}$ and $\overline{1.5}$ individually. Thus the two vectors $\overline{3^*} + \overline{1.5}$ and $\overline{1.5} + \overline{3^*}$ must have the same length, namely, $3 - 1.5$. Next consider the directions of $\overline{3^*} + \overline{1.5}$ and $\overline{1.5} + \overline{3^*}$. Again from the definition of vector addition, the direction of either sum vector is the direction of the longer vector (i.e., $\overline{3^*}$ for the case at hand, therefore left-pointing). So the two sum vectors also have the same direction. Since two vectors with the same length and the same direction are equal, we have proved that

$$\overline{3^*} + \overline{1.5} \;=\; \overline{1.5} + \overline{3^*}$$

Consider another example, $\overline{2} + \overline{1^*}$. We slide the vector $\overline{1^*}$ until its starting point (i.e., 0) is at 2, as shown:



The vector $\overline{2} + \overline{1^*}$ is therefore $\overline{1}$, the vector that starts at 0 and ends at 1. If we consider instead the sum with the order of the vectors reversed, $\overline{1^*} + \overline{2}$, then we get:



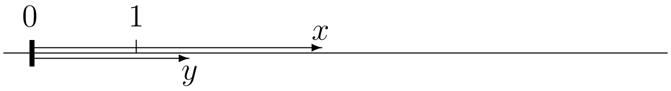
Pictorially, we see that $\overline{1^*} + \overline{2}$ is again $\overline{1}$, but we can directly prove this as follows. Because the direction of either sum, $\overline{2} + \overline{1^*}$ or $\overline{1^*} + \overline{2}$, is the direction of the longer vector (i.e., $\overline{2}$) and the length of either sum is the difference of the lengths of $\overline{2}$ and $\overline{1^*}$ (i.e., $2 - 1$), we have once more the equality:

$$\overline{2} + \overline{1^*} \;=\; \overline{1^*} + \overline{2}$$

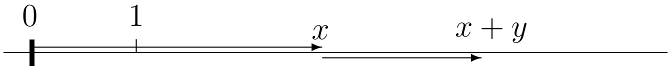
In particular, $\overline{1} = \overline{1^*} + \overline{2}$, as before.

The preceding arguments do not depend on the explicit values of $\overline{3^*}$, $\overline{1.5}$, $\overline{2}$, and $\overline{1^*}$. So the addition of vectors is commutative if the vectors point in different directions.

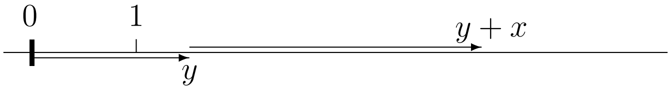
Now suppose we have two vectors $\overline{x}$ and $\overline{y}$, and both are right pointing (i.e., $x > 0$ and $y > 0$):



Here is $\overline{x} + \overline{y}$:



and here is $\overline{y} + \overline{x}$:



Because clearly the segment of either $\overline{x} + \overline{y}$ or $\overline{y} + \overline{x}$ is just the concatenation of the segments of $\overline{x}$ and $\overline{y}$, and because both $\overline{x} + \overline{y}$ and $\overline{y} + \overline{x}$ are right-pointing, we have as expected:

$$\overline{x} + \overline{y} \;=\; \overline{y} + \overline{x}$$

Of course, if both x and y are left-pointing (i.e., $x < 0$ and $y < 0$), the reasoning is entirely the same. So the addition of vectors is also commutative if the vectors point in the same direction. We have therefore proved the following theorem.

Theorem 1 *For any two numbers x and y,*

$$\overline{x} + \overline{y} \;=\; \overline{y} + \overline{x}$$

The reasoning used to prove Theorem 1 also proves the following useful result.

Theorem 2 *Let x and y be any two numbers. Then:*

(i) If both vectors $\overline{x}$ and $\overline{y}$ have the same direction, then the sum vector $\overline{x} + \overline{y}$ has the same direction as $\overline{x}$ and $\overline{y}$ and its length is the length of the concatenation of the segments of $\overline{x}$ and $\overline{y}$, and is therefore the sum of the lengths of $\overline{x}$ and $\overline{y}$.

(ii) If the vectors $\overline{x}$ and $\overline{y}$ have different directions, then the direction of the sum vector $\overline{x} + \overline{y}$ is the same as the direction of the longer vector and the length of the

sum vector $\overline{x} + \overline{y}$ is the difference of the lengths of the vectors $\overline{x}$ and $\overline{y}$.

Because a vector is completely determined by its length and its direction, Theorem 2 tells us how to add $\overline{x} + \overline{y}$ for any two numbers x and y. This will be used in the next sub-section.

Adding rational numbers

We are now in a position to define the addition of rational numbers. The **sum $\boldsymbol{x + y}$ of any two rational numbers $\boldsymbol{x}$ and $\boldsymbol{y}$** is by definition the endpoint of the vector $\overline{x} + \overline{y}$. In other words,

$$x + y = \text{ the endpoint of } \overline{x} + \overline{y}.$$

Put another way, $x + y$ is defined to be the point on the number line so that its corresponding vector $\overline{x + y}$ satisfies:

$$\boldsymbol{\overline{x + y} = \overline{x} + \overline{y}.}$$

From Theorem 1, we conclude that the addition of rational numbers is commutative. Moreover, part (i) of Theorem 2 shows that if x and y are fractions, then $x + y$ is the length of the concatenation of $[0, x]$ and $[0, y]$ and therefore has exactly the same meaning as the addition in §3 of Chapter 1. It follows that the addition of rational numbers $x + y$, as defined by vector addition, coincides with the addition of fractions in the sense of §3 of Chapter 1 when x, y are fractions. These observations are enough for the explicit determination of the sum of any two rational numbers, as follows.

A rational number is either 0, or a fraction, or the mirror reflection (across 0) of a fraction. Therefore, if s and t are two *fractions*, then

$$s + t, \quad s + t^*, \quad s^* + t, \quad \text{and} \quad s^* + t^*$$

exhaust all possibilities of the sum of two rational numbers. Knowing further that addition is commutative, we need only deal with one of $s + t^*$ and $s^* + t$ because knowing how to compute the value of either one *for arbitrary fractions s and t* means knowing how to compute the other. With this understood, the following theorem tells us how to compute the sum of any two rational numbers.

Theorem 3 *For all fractions s and t,*

$$\begin{aligned} s+t &= \textit{the ordinary sum of the fractions } s \textit{ and } t \\ s^*+t^* &= (s+t)^* \\ s+t^* &= t^*+s \;=\; \begin{cases} (s-t) & \text{if } s \geq t \\ (t-s)^* & \text{if } s < t \end{cases} \end{aligned}$$

Theorem 3 will seem much more natural once we have computed some explicit numbers. There is no need to look at $s+t$ (s and t are understood to be fractions). Consider next, $7^* + 6^*$. By definition, this is the endpoint of $\overline{7^*} + \overline{6^*}$. By part ($i$) of Theorem 2, the sum $\overline{7^*} + \overline{6^*}$ is left-pointing and has length $7+6=13$. Thus the endpoint of $\overline{7^*} + \overline{6^*}$ is $13^* = (7+6)^*$. We have proved that

$$7^* + 6^* \;=\; (7+6)^*$$

Next, we look at $10\frac{1}{2} + 3^*$. This is the endpoint of $\overline{10\frac{1}{2}} + \overline{3^*}$. By Theorem 2 ($ii$), this vector sum is right-pointing and has length $10\frac{1}{2} - 3$ and therefore its endpoint is just $10\frac{1}{2} - 3$. This we have proved that

$$10\frac{1}{2} + 3^* \;=\; 10\frac{1}{2} - 3$$

Finally, consider $11.5 + 12.1^*$. It is the endpoint of $\overline{11.5} + \overline{12.1^*}$. By part ($ii$) of Theorem 2, this vector sum is left-pointing with length equal to $12.1 - 11.5$; its endpoint is therefore $(12.1 - 11.5)^*$, and we have

$$11.5 + 12.1^* \;=\; (12.1 - 11.5)^*$$

We proceed to give a formal proof of Theorem 3.

Proof of Theorem 3 By the remarks below the definition of $x+y$, the first assertion is true. Next, consider $s^* + t^*$, which is, by definition of addition, the endpoint of the vector sum $\overline{s^*} + \overline{t^*}$. By Theorem 2($i$), the vector sum is left-pointing and its length is $s+t$; therefore the endpoint is exactly $(s+t)^*$. Thus we have proved the second assertion. Now suppose $s \geq t$ and we are given $s + t^*$. By definition of addition, this is the endpoint of the vector sum $\overline{s} + \overline{t^*}$. Since $s \geq t$, $\overline{s}$ is the longer vector. Part (ii) of Theorem 2 therefore implies that the vector sum is right-pointing and has length $s - t$. Thus the endpoint of $s + t^*$ is just $s - t$, which is the first half of the third assertion in the theorem. Finally, suppose we consider $s + t^*$ with $s < t$. This is the endpoint of the vector sum $\overline{s} + \overline{t^*}$. Since $s < t$ this time around, the longer vector is now $\overline{t^*}$ and the vector sum is left-pointing and the length is $t - s$. In other words, the endpoint of $\overline{s} + \overline{t^*}$ is $(t-s)^*$. This proves the second half of the third assertion in the theorem. The proof is complete.

Activity Compute (a) $(3\frac{7}{9})^* + (2\frac{5}{12})^*$, (b) $9.21 + (3.3)^*$, (c) $(\frac{9}{7})^* + \frac{14}{11}$.

We conclude this sub-section with a comprehensive statement about the addition of rational numbers. This will be useful for the discussion of multiplication in §4.

Theorem 4 *The addition of rational numbers satisfies the following properties:*

(A1) *It is associative and commutative, and if x and y are* fractions*, then $x + y$ is the same as the usual addition of fractions.*
(A2) $x + 0 = x$ *for any rational number x.*
(A3) *If x is any rational number, then $x + x^* = 0$.*

Proof (A2) and (A3) are immediate consequences of the definition of rational number addition. Of the assertions in (A1), the comments given immediately below the definition of the addition of rational numbers show that only the associative law remains to be proved. Thus we have to prove:

$$(x+y)+z \;=\; x+(y+z) \quad \text{for all rational numbers } x,\, y,\, z \qquad (\natural)$$

If x, y, z are all positive, this is the associativity of fraction addition, and we already know that. Suppose x, y, z are all negative. Then we appeal to the following useful lemma, already used implicitly in the proof of Theorem 3.

Lemma *For all* $x, y \in \mathbf{Q}$*,* $(x+y)^* = x^* + y^*$*.*

First let us see this Lemma is going to help with the proof of Theorem 4. For negative x, y, z, the numbers x^*, y^*, z^* are fractions and therefore

$$(x^* + y^*) + z^* \;=\; x^* + (y^* + z^*)$$

Then of course,

$$(\,(x^* + y^*) + z^*)^* \;=\; (x^* + (y^* + z^*)\,)^*$$

Now use the Lemma to conclude $(x^* + y^*)^* + z \;=\; x + (y^* + z^*)$. Using the Lemma once more, we get $(x+y)+z = x+(y+z)$, which is ($\natural$).

Let us **prove the Lemma**. Clearly the vectors $\overline{x} + \overline{y}$ and $\overline{x^*} + \overline{y^*}$ are mirror reflections of each other across 0, i.e., their endpoints $x+y$ and $x^* + y^*$ are mirror reflections of each other. But this is exactly the statement the Lemma, so the proof is complete.

Now we have to complete the proof of ($\natural$). Because of the commutativity of addition, it remains to examine ($\natural$) in the following two cases.

Case 1. Exactly one of x, y, z is negative.

Case 2. Exactly two of x, y, z are negative.

Suppose we already know that ($\natural$) is true for Case 1. We now prove that Case 2 also follows. This is because, let us say $x < 0$, $y < 0$ and $z > 0$, then the three numbers x^*, y^* and z^* satisfy Case 1 and therefore

$$(x^* + y^*) + z^* \;=\; x^* + (y^* + z^*)$$

As before, if we take the * of both sides and apply the Lemma, we get $(x+y)+z \;=\; x+(y+z)$, and ($\natural$) holds also for Case 2.

It remains to prove ($\natural$) for Case 1. Before we embark on the proof, let us test it with some specific numbers. Let us show:

$$(17^* + 5) + 9 \;=\; 17^* + (5+9)$$

Direct computations using Theorem 3 show that the left side is $12^* + 9 = 3^*$, whereas the right side is $17^* + 14 = 3^*$. So they are equal. Can we see any reason behind the computation? Yes, because if we use Theorem 2, then we see that $(17^* + 5)$ is the endpoint of a vector that is left-pointing with length $17 - 5 = 12$, so that $(17^* + 5) + 9$ is the endpoint of another vector which is still left-pointing but now has length $(17 - 5) - 9 = 17 - 5 - 9$. As to the right side, $(5 + 9)$ is straightforward as it is the endpoint of a right-pointing vector with length 14, so that $17^* + (5 + 9)$ is the endpoint of a left-pointing vector of length $17 - 14 = 17 - (5 + 9) = 17 - 5 - 9$. So both sides are the endpoint of the same vector and must therefore be equal.

Let us try something else by proving

$$(17^* + 5) + 15 = 17^* + (5 + 15)$$

Again, direct computations using Theorem 3 show that both sides are equal to 3. We can also look at each side in terms of vectors and make use of Theorem 2. As we have seen, $(17^* + 5)$ is the endpoint of a left-pointing vector with length $17 - 5$, so that $(17^* + 5) + 15$ is the endpoint of a right-pointing vector with length $15 - (17 - 5) = 15 - 17 + 5$. The right side $17^* + (5 + 9)$ is the endpoint of a right-pointing vector with length $(5 + 15) - 17 = 5 + 15 - 17$. Once again, the two sides are equal.

We note that the proof for a different triple of numbers,

$$(17^* + 25) + 15 = 17^* + (25 + 15),$$

would be qualitatively different from the preceding ones because all the vector sums associated with the additions — $(17^* + 25)$, $(17^* + 25) + 15$, $(25 + 15)$, and $17^* + (25 + 15)$ — are right-pointing. The absence of a uniform pattern therefore presages the need of a case by case analysis in the general proof, and this turns out to be correct. *The following proof is long and tedious, and should be skipped on first reading.*

We may assume $x > 0$, $y > 0$ but $z < 0$, as the other possibilities are similar. Thus we must prove:

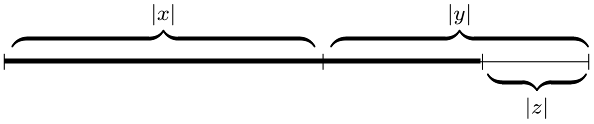
$$(\overline{x} + \overline{y}) + \overline{z} \;=\; \overline{x} + (\overline{y} + \overline{z}) \quad \text{where } x > 0,\ y > 0 \text{ but } z < 0 \qquad (\natural\natural)$$

We will split this proof into yet another two cases. Let us denote the length of a vector $\overline{x}$ by $|\boldsymbol{x}|$.[32] Then we have

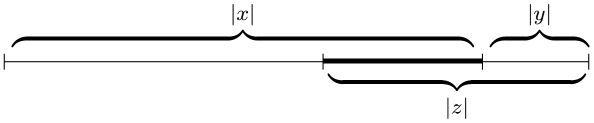
[32] If this notation reminds you of *absolute value* (to be defined later in §6 of this chapter), rest assured that it is intentional.

$$\textit{Case A: } |\overline{x}+\overline{y}| \geq |\overline{z}|, \quad \text{and} \quad \textit{Case B: } |\overline{x}+\overline{y}| < |\overline{z}|.$$

Case A. By Theorem 2, the left side of $(\natural\natural)$ is a right-pointing vector of length $(|x|+|y|)-|z|$. The right side of $(\natural\natural)$ is $\overline{x}+(\overline{y}+\overline{z})$. *Assume that* $|y| \geq |z|$. Then by Theorem 2, $\overline{y}+\overline{z}$ is right-pointing of length $|y|-|z|$ or 0. Therefore by part (i) of Theorem 2, $\overline{x}+(\overline{y}+\overline{z})$ is right-pointing with length $|x|+(|y|-|z|)$, which is equal to $(|x|+|y|)-|z|$ by the definition of subtraction, as the following picture shows:

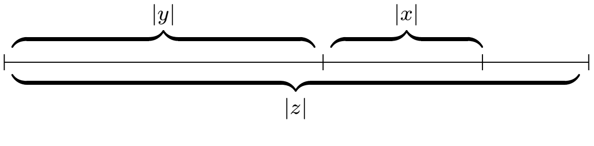


So $(\natural\natural)$ is proved for this situation. On the other hand, if instead we *assume* $|y| < |z|$, then by Theorem 2 again, $\overline{y}+\overline{z}$ is left pointing of length $|z|-|y|$. We claim that $\overline{x}$ is longer than $\overline{y}+\overline{z}$. To prove the claim, recall that we are assuming $|\overline{x}+\overline{y}| \geq |\overline{z}|$, i.e., $|x|+|y| > |z|$. Therefore $|x| > |z|-|y|$, as the following picture shows (the thickened segment has length $|z|-|y|$):



Thus $\overline{x}$ is longer than $\overline{y}+\overline{z}$, so that part (ii) of Theorem 2 shows $\overline{x}+(\overline{y}+\overline{z})$ is right-pointing and has length $|x|-(|z|-|y|)$. From the definition of subtraction, the preceding picture also shows $|x|-(|z|-|y|) = (|x|+|y|)-|z|$, so $(\natural\natural)$ is completely proved for Case A..

Case B. We now assume $|\overline{x}+\overline{y}| < |\overline{z}|$. The left side of $(\natural\natural)$, which is $(\overline{x}+\overline{y})+\overline{z}$, is now a left-pointing vector of length $|z|-(|x|+|y|)$. Now we look at the right side $\overline{x}+(\overline{y}+\overline{z})$. The inequality $|\overline{x}+\overline{y}| < |\overline{z}|$ implies in particular that $|y| < |z|$. Theorem 2 therefore implies that $\overline{y}+\overline{z}$ is a left-pointing vector of length $|z|-|y|$. Now, using the fact that $|\overline{x}+\overline{y}| < |\overline{z}|$, we have the following picture which shows that $|x| < |z|-|y|$.



The inequality $|x| < |z|-|y|$ shows that the vector $\overline{x}+(\overline{y}+\overline{z})$ is left-pointing and its length is $(|z|-|y|)-|x|)$ which, as the preceding picture shows, is equal to

$|z| - (|x| + |y|)$. Thus $\overline{x} + (\overline{y} + \overline{z})$ is the same vector as $(\overline{x} + \overline{y}) + \overline{z}$, and the proof of Case B, and therewith the proof of Theorem 4, is complete.

Exercises 2.2

1. Prove that or all $x, y \in \mathbf{Q}$, if $x + y = x$, then $y = 0$.

2. For each of the following numbers, explain as if to a seventh grader whether it is positive or negative:

$$(68\tfrac{1}{2})^* + 68\tfrac{2}{5}, \quad (1\tfrac{7}{8})^* + 2\tfrac{1}{10}, \quad \tfrac{16}{7} + (2\tfrac{1}{4})^*, \quad (1\tfrac{3}{10})^* + \tfrac{9}{7}.$$

3. $(\frac{87}{89})^* + \left((104\frac{10}{117})^* + (\frac{2}{89})^*\right) + \left(10 + (105\frac{10}{117})^*\right)^* = ?$

4. Explain as if to a seventh grader, directly and *without* making use of Theorem 3, why $(2.3)^* + (1\frac{2}{5})^* = 3.7^*$, and why $(9\frac{1}{2})^* + 7.5 = 2^*$.

5. Compute: (a) $(4\frac{6}{7})^* + 2\frac{2}{3}$. (b) $7.1^* + (22\frac{1}{3})^*$. (c) $(4\frac{2}{101})^* + (2.5 + 3\frac{99}{101})^*$. (d) $(703.2^* + 689.4) + \left(\frac{1}{5} + 3\frac{2}{3}\right)^*$. (e) $\left(\frac{5}{6} + \left(1\frac{7}{18}\right)^*\right) + \frac{5}{24}$.

6. Give a direct proof of the associative law of addition in the following two special cases: $(3 + 6.5^*) + 2.5 = 3 + (6.5^* + 2.5)$, and $(2.5^* + 1.8^*) + 3.7 = 2.5^* + (1.8^* + 3.7)$.

7. For all $x, y \in \mathbf{Q}$, if $x + y = 0$, prove that $y = x^*$ and $x = y^*$.

3 Subtracting rational numbers

Subtraction as addition

The introduction of $-x$

Subtraction as addition

The explicit formulas for the addition of rational numbers in Theorem 3 lead to the following insight: *the subtraction of fractions becomes addition in the larger context of rational numbers,* in the following sense. If s, t are fractions so that $s \geq t$, then $s - t$ makes sense as ordinary fraction subtraction (see §3 of Chapter 1) while $s + t^*$ also makes sense as the *addition* of two rational numbers; Theorem 3 affirms that both give the same result. The fact that, although $s - t$ up to this point has no meaning when $s < t$, $s + t^*$ makes sense for all $s, t \in \mathbf{Q}$ immediately suggests that we could *define* in general the subtraction between any two *rational numbers* s and t to be just $s + t*$. Formally, for rational numbers x and y, we define the **subtraction** $\boldsymbol{x - y}$ **as**

$$\boldsymbol{x - y} \quad \overset{\textbf{def}}{=} \quad \boldsymbol{x + y^*}$$

We emphasize the idea, already mentioned several times, that this concept of subtraction between rational numbers $x - y$ coincides with the previous concept of fraction subtraction when x and y are fractions. We build on what we know rather than play a new game at every turn. Thus, a subtraction such as $\frac{6}{5} - \frac{1}{4}$ has exactly the same meaning whether we look at it as a subtraction between the two fractions $\frac{6}{5}$ and $\frac{1}{4}$ *or* between these fractions considered as rational numbers. On the other hand, we are now free to do a subtraction between any two fractions such as $\frac{1}{4} - \frac{6}{5}$ even when the first fraction is smaller than the second. We see for the first time the advantage of having rational numbers available: we can as freely subtract any two fractions as we add them. But this goes further, because we can even subtract not just any two fractions, but any two *rational numbers*, e.g., $5.5^* - 17^*$.

This definition reveals that *subtraction is just a different way of writing addition* among rational numbers. Hence any property about subtraction among rational numbers is ultimately one about addition.

In the rest of the section, we will explore a bit the ramifications of this concept of subtraction. The overriding fact is that, without this general definition, we do not have a good grasp of what subtraction is about. Beyond the oddity of not being able to subtract a larger fraction from a smaller one, there is also the unpleasant observation that "subtraction is not associative", i.e., in general, $(x-y)-z \neq x-(y-z)$ for fractions x, y, z. For example, letting $x=4$, $y=2$, $z=1$, the left side is 1 while the right side is 3. We proceed to clarity this situation.

The introduction of $-x$

We start from the beginning. As a consequence of the definition of $x-y$, we have

$$0-y=y^*$$

because $0-y=0+y^*=y^*$ (see (A2) of Theorem 4). Common sense dictates that we should abbreviate $0-y$ to $-y$, so that we write from now on:

$$\boldsymbol{-y \;=\; y^*}$$

At this point, we abandon the notation of y^* and replace it by the more common $-y$. We call $-y$ **minus $\boldsymbol{y}$**.

Let us go through a few expressions involving y^* and see how, *strictly according to the definitions*, we can smoothly transition to the new notation. This is of some importance because we hope this is how you are going to teach your students and impress on them that, in mathematics, everything proceeds according to reason and nothing is done on account of somebody's whims.

For any $x \in \mathbf{Q}$, we have $x+(-y)=x+y^*=x-y$, where the last equality is by definition of $x-y$. Therefore

$$\boldsymbol{x+(-y) \;=\; x-y} \quad \text{for all } \boldsymbol{x, y} \in \mathbf{Q}$$

Next we show

$$\boldsymbol{-x+y \;=\; y-x} \quad \text{for all } \boldsymbol{x, y} \in \mathbf{Q}$$

This is because

$$\begin{aligned} y-x &= y+x^* && \text{(definition of subtraction)} \\ &= x^*+y && \text{(commutativity of addition)} \\ &= -x+y \end{aligned}$$

For example, $-\frac{2}{3}+4=4-\frac{2}{3}$, both being equal to $\frac{10}{3}$ as a simple application of Theorem 3 shows.

Letting y be $-y$ in the preceding equality then gives the "*commutativity of subtraction*":

$$-x - y \;=\; -y - x \quad \text{for all } x, y \in \mathbf{Q}$$

Let us re-state two of our previous conclusions in the new notation. From $x^{**} = x$ for any $x \in \mathbf{Q}$, we get

$$-(-x) = x$$

(A3) of Theorem 4 now states that

> **(A3*)** *If x is any rational number, $x + (-x) = (-x) + x = 0$.*

The Lemma and Theorems 3 now read:

Lemma* *For all $x, y \in \mathbf{Q}$, $-(x+y) \;=\; -x - y$.*

(For Lemma*, observe that $x^* + y^* = x^* - y$, by the definition of subtraction, so that $x^* + y^* = -x - y$.)

Theorem 3* *For all* fractions *s and t,*

$$\begin{aligned} s + t \;&=\; \textit{the ordinary sum of the fractions } s \textit{ and } t \\ -s - t \;&=\; -\,(s+t) \\ s - t \;&=\; -\,t + s \;=\; \begin{cases} (s-t) & \text{if } s \geq t \\ -\,(t-s) & \text{if } s < t \end{cases} \end{aligned}$$

Observation: We now see that Lemma*, in the form of

$$-(x+y) \;=\; -x - y$$

for all rational numbers x and y, *is* a statement about "removing parentheses". We can go a step further: for all rational numbers x and y,

$$-(x-y) \;=\; -x + y \qquad \text{and} \qquad -\,(-x+y) = x - y$$

We leave these as exercises.

We pursue the theme that subtraction is another way of writing addition among rational numbers and bring closure to a remark we made at the end of §3 in Chapter 1 about the subtraction of fractions. We now show that for any rational numbers a, b, x, y,

$$(a+b)-(x+y)=(a-x)+(b-y)$$

This is because

$$(a+b)-(x+y)=a+b+(x+y)^*=a+b+x^*+y^*$$

where the first equality is by the definition of subtraction and the second equality is on account of the Lemma. Thus $(a+b)-(x+y)=(a+x^*)+(b+y^*)$, by Theorem 1 of the Appendix in Chapter 1. By the definition of subtraction again, we get $(a+b)-(x+y)=(a-x)+(b-y)$.

It is clear from this reasoning that there is a similar assertion if $a+b$ is replaced by a sum of k rational numbers for any positive integer k and the same is done to $x+y$. The details are left as an exercise.

Finally we take up the issue of why, on the basis of the associative law of addition, $(x-y)-z \neq x-(y-z)$. We have:

$$\begin{aligned}(x-y)-z &= (x+y^*)+z^* && \text{(definition of subtraction)}\\ &= x+(y^*+z^*) && \text{(associativity)}\\ &= x+(y+z)^* && \text{(Theorem 1)}\\ &= x-(y+z) && \text{(definition of subtraction)}\end{aligned}$$

Therefore

$$(x-y)-z \;=\; x-(y+z)$$

and this is why $(x-y)-z \neq x-(y-z)$.

Exercises 2.3

1. Without using Theorem 3 or Theorem 3*, and using only Theorem 2, explain as if to a seventh grader why $\frac{4}{3}-2\frac{1}{5}=-\frac{13}{15}$.

2. Prove that for all rational numbers x and y, we have $-(x-y) = -x+y$ and $-(-x+y) = x-y$. Give the reason at each step.

3. Explain carefully why each of the following is true for all rational numbers x, y, z: (a) $(x+y)-z = x+(y-z)$. (b) $(x-y)-z = (x-z)-y$. (c) $x-(y-z) = (x-y)+z$.

4. (a) Explain as if to seventh graders why $-7584\frac{1}{279} = -7584 - \frac{1}{279}$. (b) Explain as if to seventh graders in two different ways why $-7584\frac{1}{279} < -7584$. (c) Explain as if to seventh graders why it is not a good idea to do (a) or (b) by converting the mixed number to a fraction first.

5. Compute and explain every step: $-97654\frac{1}{123} - \left(\frac{122}{123\times 124} - 97644\frac{1}{124}\right)$.

6. Compute: (a) $(-5\frac{2}{5}) - (-6)$. (b) $(-\frac{7}{12}) - (-\frac{12}{21})$. (c) $5\frac{1}{2} - (\frac{27}{5} - 5\frac{13}{15})$.

7. (a) Let a, b, $\ldots$, z, w be rational numbers. Give a detailed proof of

$$(a+b+c+d) - (x+y+z+w) = (a-x) + (b-y) + (c-z) + (d-w)$$

by justifying every step. (b) Can you extend (a) from a pair of 4 rational numbers to a pair of n rational numbers for any positive integer n? For notation, try

$$(a_1 + a_2 + \cdots + a_n) - (x_1 + x_2 + \cdots + x_n) = (a_1 - x_1) + (a_2 - x_2) + \cdots + (a_n - x_n).$$

4 Multiplying rational numbers

Assumptions on multiplication

We now take up the multiplication of rational numbers. If we imitate what we did with addition by giving a general definition of the product of two rational numbers and then show that the definition yields the same product as in §4 of Chapter 1 in case the rational numbers are fractions, then we will discover that the definition consists of the following four rules: for all fractions s and t,

$$\begin{aligned} st &= \text{same product } st \text{ as before} \\ (-s)t &= -(st) \\ s(-t) &= -(st) \\ (-s)(-t) &= st \end{aligned}$$

This was pretty much what Diophantus[33] did when he first introduced negative numbers. One can infer from his writing why he defined multiplication this way, and we could very well just follow Diophantus' brute force method and say, "*There it is.*" However, it may be more enlightening to *try to* retrace Diophantus' steps in order to see for ourselves *why* things are the way they are. In other words, imagine that we have gotten used to working with positive numbers and we are suddenly confronted with the need to deal with negative numbers (which was pretty much what Diophantus had to face). For example, if we try to write down a solution to $x + 7 = 1$, then we would have a negative number staring us in the face and we would have to compute with them, willy nilly (see the discussion at the beginning of §2). In particular, we have to learn to multiply such numbers, and we do so without doing violence to the existing rules of arithmetic. Of course we'd take for granted that there *is* a way to multiply them; that is not in doubt! Moreover,*to judge by the usual way of computing with positive numbers, we expect that multiplication will continue to be associative, commutative, and distributive, and that when a negative number is multiplied by 1 it doesn't change.* If we are going to hold onto these beliefs, then we will be forced to multiply rational numbers in the way described above. What we are going to do now

[33] A Greek mathematician who lived in Alexandria, Egypt, *probably* around 250 AD. Incidentally, the female mathematician Hypatia (c. 355-415 AD) also spent her life in Alexandria; her murder, engineered by St. Cyril of Alexandria, brought to a close the classical period of Greek mathematics.

is give the details of why this is so.

Let us then be clear about what we are going to take for granted for the rest of this section. We will summarize it in the form of the following two **fundamental assumptions on multiplication**:

> **(M1)** *Given any two rational numbers x and y, there is a way to multiply them to get another rational number xy so that, if x and y are fractions, xy is the usual product of fractions. Furthermore, this multiplication of rational numbers satisfies the associative, commutative, and* distributive *laws.*
>
> **(M2)** *If x is any rational number, then* $1 \cdot x = x$.

We note that (M2) says, in particular, that $1 \times (-5.4) = -5.4$, and this fact is certainly new.

Now assumptions (M1) and (M2) remind us of (A1) and (A2) in Theorem 4 of §2, with 1 playing the role in multiplication what 0 does in addition. There is no analog of (A3) at the moment because the counterpart of x^* in addition is the number x^{-1} in multiplication, and the latter will not be defined until §5.

On the basis of (M1) and (M2), we are going to learn how rational numbers *must be multiplied.* There is an "obvious" fact that we can dispose of right away.

Lemma 1 $0 \cdot x = 0$ *for any* $x \in \mathbf{Q}$.

Proof We know from the definition of addition that $0 + 0 = 0$. Take an $x \in \mathbf{Q}$, then $0 \cdot x = (0+0) \cdot x = 0 \cdot x + 0 \cdot x$, by the distributive law. Thus $0 \cdot x = 0 \cdot x + 0 \cdot x$, so that

$$0 \cdot x - 0 \cdot x \;=\; (0 \cdot x + 0 \cdot x) - 0 \cdot x.$$

Of course the left side is equal to 0. The right side is $(0 \cdot x + 0 \cdot x) + (0 \cdot x)^*$ so that, by the associative law, it becomes

$$0 \cdot x + (0 \cdot x + (0 \cdot x)^*) \;=\; 0 \cdot x + 0 \;=\; 0 \cdot x$$

Altogether, we get $0 = 0 \cdot x$, and Lemma 1 is proved.

The equality $(-m)(-n) = mn$ for whole numbers

Once we have introduced multiplication among rational numbers, our first task is to find out how multiplication is related to the existing operations, in particular, addition and the mirror reflection $*$. As always, the relationship between addition and multiplication is codified by the distributive law, which we must point out is part of the assumption in (M1). As to the operation $*$, we can ask whether the order of applying multiplication and $*$ is interchangeable. In other words, given two rational numbers x and y, if we get their mirror reflections first and then multiply (thus x^*y^*), how is it related to the number obtained by multiplying them first and then getting its mirror reflection (thus $(xy)^*$). If multiplication is replaced by addition, the order is interchangeable; see the Lemma of §2. *In the case of multiplication, however, the order matters.* In fact, $x^*y^* \neq (xy)^*$, or in the notation of the minus sign, $(-x)(-y) \neq -(xy)$. As is well-known, the correct answer is

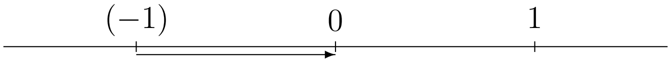
$$(-x)(-y) = xy \qquad \text{for all rational numbers } x \text{ and } y.$$

This surprising fact, the bane of many middle school students, can be given a very short proof. We will present this proof at the end of the next sub-section. For the middle school classroom, such a proof is too sophisticated to be given at the outset (or maybe never in school mathematics). Instead we will give a more leisurely proof by first taking a detour through the more familiar terrain of the integers to see why $(-m)(-n) = mn$ for all *whole numbers* m and n. There is a reason for singling out the whole numbers. It is not only easier to learn (it is that for sure!), but it is also far easier to *teach*, as we shall see presently. If you can get *all* your kids to believe, for example, that $(-1234)(-5678) = 1234 \times 5678$, then you are already ahead of the game.

We begin with the simplest special case of this assertion: the case of $x = y = 1$. This will turn out to be the critical case.

Theorem 1 $(-1)(-1) = 1$.

Proof Let x denote $(-1)(-1)$. Our goal is to show $x = 1$. We should ask ourselves: if we have a number x, how can we tell if it is 1 or not? One way is to try to see if $(-1)+x = 0$. If it is, then $\overline{x}$ is the right-pointing vector of length 1 (because it must go from -1 back to 0), and therefore $x = 1$.[34]



With this in mind, we now compute:

$$\begin{aligned} (-1)+x &= 1\cdot(-1)+(-1)(-1) && \text{(by (M2) and the definition of } x\text{)} \\ &= (1+(-1))\cdot(-1) && \text{(distributive law)} \\ &= 0\cdot(-1) \\ &= 0 && \text{(Lemma 1)} \end{aligned}$$

So we know $x = 1$, i.e., $(-1)(-1) = 1$. The proof is complete.

We can rephrase this proof to give it a more algebraic flavor; this reformulation will turn out to be more in line with the proof in general that $(-x)(-y) = xy$ for all $x, y \in \mathbf{Q}$. It goes as follows:

We have $(-1)+1 = 0$ (review (A3) of Theorem 4 in §2 if necessary). Multiply each side by (-1) and apply the distributive law to get

$$(-1)(-1) + 1\cdot(-1) \;=\; 0\cdot(-1)$$

By Lemma 1, the right side is 0. As to the left side, by (M2), it is equal to $(-1)(-1) + (-1)$. Therefore,

$$(-1)(-1) + (-1) \;=\; 0$$

Adding 1 to both sides and using the associative law of addition, we get:

$$(-1)(-1) + ((-1)+1) \;=\; 0+1$$

Now the left side is $(-1)(-1)$ and the right side is 1. Theorem 1 is proved.

[34] This is entirely similar to what one does in chemistry: to see whether a solution is acidic or not, dip a piece of litmus paper (ph test strip) into the solution and see if it turns red.

Activity Practice explaining as if to a seventh grader why $(-1)(-1) = 1$ by using your neighbor as a stand-in for the seventh grader. (Theorem 1 is so basic to understanding rational number multiplication that this Activity is strongly recommended.)

We can now give the proof that $(-m)(-n) = mn$ for all *whole numbers m and n*. Let us first do a special case: why is $(-2)(-3) = 6$? This is because

$$\begin{aligned}
(-2)(-3) &= \{(-1)+(-1)\}\cdot\{(-1)+(-1)+(-1)\} && \text{(Theorem 3* of §3)}\\
&= \underbrace{(-1)(-1)+\cdots+(-1)(-1)}_{6} && \text{(distributive law)}\\
&= \underbrace{1+\cdots+1}_{6} && \text{(Theorem 1)}\\
&= 6
\end{aligned}$$

In the same manner, we can show even fifth graders why $(-3)(-4) = 12$, $(-5)(-2) = 10$, etc. A teacher can probably win the psychological battle over students' disbelief of the "(negative)×(negative) = positive" phenomenon by these very concrete computations.

The general proof of $(-m)(-n) = mn$ is essentially the same. So let m, n be whole numbers. We first prove that

$$(-1)(-m) \;=\; m$$

This is because

$$\begin{aligned}
(-1)(-m) &= (-1)\{\underbrace{(-1)+\cdots+(-1)}_{m}\} && \text{(Theorem 3* of §3)}\\
&= \underbrace{(-1)(-1)+\cdots+(-1)(-1)}_{m} && \text{(distributive law)}\\
&= \underbrace{1+\cdots+1}_{m} && \text{(Theorem 1)}\\
&= m
\end{aligned}$$

Hence, we have:

$$\begin{aligned}(-n)(-m) &= (\underbrace{(-1)+\cdots+(-1)}_{n})(-m) && \text{(Theorem 3* of §3)}\\ &= \underbrace{(-1)(-m)+\cdots+(-1)(-m)}_{n} && \text{(distributive law)}\\ &= \underbrace{m+\cdots+m}_{n}\\ &= nm \;=\; mn\end{aligned}$$

Multiplication of rational numbers in general

Our goal is to find out explicitly how to multiply rational numbers. As noted, since a nonzero rational number is either a fraction or a negative fraction, it is a matter of finding out the values of the following four products where s and t are nonzero fractions:

$$st, \quad (-s)t, \quad s(-t), \quad \text{and} \quad (-s)(-t).$$

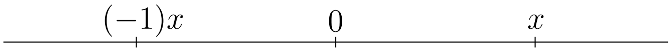
In the last sub-section, we already got a taste of what to expect if s and t are whole numbers. Therefore we can afford to directly attack a more general problem, that of determining the values of

$$xy, \quad (-x)y, \quad x(-y), \quad \text{and} \quad (-x)(-y),.$$

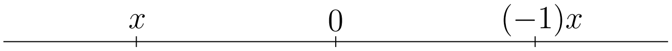
where x and y are arbitrary *rational numbers*. We first prove the following generalization of Theorem 1.

Theorem 2 *For any rational number x, the number $(-1)x$ is the mirror reflection of x. In symbols: $(-1)x = -x$.*

Proof The number $-x$ is the point on the opposite side of 0 from x so that x and $-x$ are equi-distant from 0. Therefore this is the picture we want to be true when x is positive:



and this is the picture we want to be true when x is negative:



Now think of the sum $x+(-1)x$ in terms of vectors (see §2). If we can show that

$$x+(-1)x \;=\; 0$$

then the vectors $\overline{x}$ and $\overline{(-1)x}$ must have opposite direction and equal length. Consequently, $(-1)x$ will have to be equal to $-x$. Let us therefore prove that $x+(-1)x$ is equal to 0. We use the distributive law:

$$x+(-1)x \overset{\text{(M2)}}{=} 1\cdot x+(-1)x \;=\; \{1+(-1)\}\,x$$

But $1+(-1)=0$. Therefore,

$$x+(-1)x = \{1+(-1)\}\,x = 0\cdot x = 0$$

where the last equality is due to Lemma 1. Theorem 2 is proved.

Remark The most critical step of the preceding proof of Theorem 2 is to convert $x+(-1)x$ to $\{1+(-1)\}x$ so that we get to $0\cdot x=0$. This is the distributive law at work. Refer back to Diophantus' initial confrontation with negative numbers some eighteen centuries ago (see the beginning of this section), we can imagine him pondering over the product $(-7)5$ and wondered what it should be. He realized that $(-7)5+7\cdot 5=((-7)+7)5$ because the distributive law[35] "must" hold. Consequently $(-7)5+7\cdot 5=((-7)+7)5=0\cdot 5=0$, so naturally he would guess $(-7)5=-(7\cdot 5)$. Since there is nothing special about the numbers 7 and 5, he would also guess that $(-x)y$ must be equal to $-(xy)$ for any numbers x and y. It is in this sense that *the distributive law "forces" the rule that* $(-x)y=-(xy)$. We can pursue this argument to see how the other rules of Diophantus must also follow from the assumed validity of commutativity, associativity and distributivity. You will see these arguments below.

[35] The explicit recognition of this concept of distributivity came only in the twelfth century, but Diophantus would have taken it for granted.

We are now ready for the general statements about rational number multiplication.

Theorem 3 *For all rational numbers x and y,*

$$\begin{aligned}(-x)y \;=\; x(-y) &\;=\; -(xy)\\ (-x)(-y) &\;=\; xy\end{aligned}$$

Proof We first prove $(-x)y = x(-y) = -(xy)$. If we read the equality of Theorem 2 backward, we get $-x = (-1)x$. Therefore $(-x)y = ((-1)x)\,y = (-1)\,(xy)$, by the associative law of multiplication. Now we apply Theorem 2 again in the same way, but this time to the rational number (xy). Then we get $(-1)(xy) = -(xy)$. Hence

$$(-x)y = (-1)(xy) = -(xy)$$

The proof of $x(-y) = -(xy)$ is similar (or we can apply the commutative law twice to what we have just proved: $x(-y) = (-y)x = -(yx) = -(xy)$).

Next, we prove $(-x)(-y) = xy$. Theorem 2 gives $(-x)(-y) = (-1)x \cdot (-1)y$. By the commutative law of multiplication (see Theorem 2 in the Appendix of Chapter 1), we have:

$$(-1)x \cdot (-1)y \;=\; (-1)(-1)(xy)$$

So the Corollary to Theorem 2 says $(-1)(-1)(xy) = 1 \cdot (xy) = xy$. Theorem 3 is proved.

Finally, by letting x and y be fractions in Theorem 3, and taking into account of (M1), we recover the rules of Diophantus:

Corollary *For all fractions s and t,*

$$\begin{aligned}st &\;=\; \textit{the ordinary product of the fractions } s \textit{ and } t\\ (-s)t &\;=\; -(st)\\ s(-t) &\;=\; -(st)\\ (-s)(-t) &\;=\; st\end{aligned}$$

Looking ahead to algebra, this Corollary is the reason that, for a number x, an expression such as $-3x$ is completely unambiguous. Indeed, this expression can be interpreted as either $-(3x)$ or $(-3)x$, but the Corollary says that both numbers are equal.

From the Corollary, we deduce the following well-known rules:

$$\begin{array}{rcl} \text{positive} \times \text{positive} & = & \text{positive} \\ \text{positive} \times \text{negative} & = & \text{negative} \\ \text{negative} \times \text{negative} & = & \text{positive} \end{array}$$

In particular, we know that

$$x^2 \geq 0 \quad \text{for any } x \in \mathbf{Q}$$

regardless of whether it is 0 or positive or negative. By FASM, we have

$$x^2 \geq 0 \quad \text{for any number } x$$

It remains to bring closure to this discussion of multiplication by delivering on a promise made at the beginning of the last sub-section, to the effect that there is a short and self-contained proof of Theorem 3.

We first prove $(-x)y = -(xy)$, where $x, y \in \mathbf{Q}$. It suffices to prove that $(-x)y + xy = 0$ (compare problem 7 in Exercises 2.2). This is so because by the distributive law,

$$(-x)y + xy = ((-x) + x)\,y = 0 \cdot y = 0$$

Next we prove $(-x)(-y) = xy$. Again, it suffices to prove that $(-x)(-y) + (-(xy)) = 0$, because this equality implies that $(-x)(-y)$ is equal to $-(-(xy))$, which is xy. Now

$$\begin{array}{rcll} (-x)(-y) + (-(xy)) & = & (-x)(-y) + ((-x)\,y) & \text{(because } -(xy) = (-x)y) \\ & = & (-x)\,((-y) + y) & \text{(distributive law)} \\ & = & (-x) \cdot 0 \;=\; 0 & \end{array}$$

The proof of Theorem 3 is complete.

Miscellaneous remarks

We conclude this section with three remarks. First, there is a simple consequence of Theorem 3 which is an explicit algorithm for the multiplication of rational numbers: if $\frac{m}{n}$ and $\frac{k}{\ell}$ are fractions, then:

$$\begin{aligned} \frac{m}{n} \times -\frac{k}{\ell} &= -\frac{mk}{n\ell} \\ -\frac{m}{n} \times \frac{k}{\ell} &= -\frac{mk}{n\ell} \\ -\frac{m}{n} \times -\frac{k}{\ell} &= \frac{mk}{n\ell} \end{aligned}$$

In the next section, we will see that these formulas remain valid even when m, n, k, ℓ are rational numbers (rather than just whole numbers).

Second, Theorem 2 gives us another way to *think of* how to remove parentheses, to the effect that $-(x+y) = -x - y$ for all $x, y \in \mathbf{Q}$ (see Lemma* of §3). This is because $-(x+y) = (-1)(x+y)$, by Theorem 2. Applying the distributive law, we get $-(x+y) = (-1)x + (-1)y = -x - y$, and the last step is by Theorem 2 again.

Third, we use Theorem 3 to tie up a loose end by proving the following form of the **distributive law for subtraction**, which is commonly taken for granted:

$$x(y-z) = xy - xz \qquad \text{for all } x, y, z \in \mathbf{Q}$$

Indeed, by using the ordinary distributive law, we have: $x(y-z) = x(y+z^*) = xy + xz^* = xy + x(-z)$. But $xy + x(-z) = xy + (-xz)$, by Theorem 3, so $x(y-z) = xy + (-xz)$. As noted in §3, we have $xy + (-xz) = xy - xz$, so the distributive law is completely proved.

It remains to supplement the second remark by pointing out that, for school students, it is undoubtedly easier to *think of* $-(x+y) = -x-y$ in terms of multiplication by -1 and the distributive law. Nevertheless, it is good to keep in mind that this equality is not about either, because it is an elementary statement *about addition* *only* and not about the far more subtle concept of rational number multiplication. So don't lose sight of the original proof of the Lemma in §2.

Exercises 2.4

1. Show that Theorem 1 follows from Theorem 2.

2. Compute, and justify each step: (a) $(-4)(-1\frac{1}{2}+\frac{1}{4})$. (b) $165 - 560(\frac{3}{4}-\frac{8}{7})$. (c) $(-\frac{3}{2})(0.64-\frac{4}{3})$. (d) $(20\frac{2}{9}\times(-\frac{5}{17}))+(3\frac{2}{9}\times\frac{5}{17})$.

3. Write out a direct explanation, in language as simple as possible, for a seventh grader of why $(-3)(-4)=3\times 4$.

4. Give as simple a proof as you can, without using Theorem 3, that, for all *whole numbers* m and n, $(-m)n=-(mn)$.

5. Use Theorem 3 to prove the other two rules of "removing parentheses":

$$-(x-y) \;=\; -x+y \qquad \text{and} \qquad -(-x+y)=x-y$$

for all rational numbers x and y. (To give proper perspective to this problem, see the concluding remark of this section.)

6. Consider each of the following two statements about any rational number x:

(a) $3x < x$.

(b) $\frac{1}{10}x > x$.

If it is always true or always false, prove. If it is sometimes true and sometimes false, give examples to explain why.

7. The following is a standard argument in textbooks to show, for example, that $(-2)(-3)=6$:

> Consider the sequence of products
>
> $$\begin{array}{llll} \cdots\cdots & 4\times(-3)=-12, & 3\times(-3)=-9, & 2\times(-3)=-6, \\ 1\times(-3)=-3, & 0\times(-3)=0, & (-1)(-3)=a, & (-2)(-3)=b, \\ (-3)(-3)=c, & (-4)(-3)=d, & \cdots\cdots & \end{array}$$
>
> Observe the pattern that, for $m\times(-3)$ as m decreases to 0, each product increases by 3. To continue this pattern beyond 0, one should assign 3 to a, 6 to b, 9 to c, 12 to d, and so on, because $(-1)(-3)=0+3=3$, $(-2)(-3)=3+3=6$, $(-3)(-3)=6+3=9$, $(-4)(-3)=9+3=12$.

Is this a valid argument? What are the implicit assumptions used? Write a critique. (*Hint:* If you write down precisely what this so-called pattern says, it would be the statement that $(n-1)(-3)=n(-3)+3$ for any positive integer n.)

8. (a) I have a rational number x so that $5-(2x-1)=(1-\frac{8}{3}x)$. What is this x? (b) Same question for $(2-3x)-(x+1)=\frac{5}{3}x+\frac{1}{2}$.

9. [For this problem, we extend the definition in §2 of Chapter 1 by defining, for any *rational number* y and any fraction $\frac{m}{n}$, the meaning of $\boldsymbol{\frac{m}{n}}$ **of** $\boldsymbol{y}$ to be $\frac{m}{n}\times y$.] (a) A rational number y has the property that $\frac{3}{4}$ of $-y$ exceeds y itself by 49. What is this y? (b) A number t has the property that twice t exceeds t^2 by $\frac{4}{7}$ of t. Find t.

5 Dividing rational numbers

Definition of division

Rational number as division of integers

Rational quotients

Definition of division

The concept of the division of rational numbers is the same as that of dividing whole numbers or dividing fractions. See the first sub-section in §5 of Chapter 1.

We begin such a discussion with the proof of a theorem that is the counterpart of Theorem 1 in §5 of Chapter 1.

Theorem 1 *Given rational numbers x and y, with $y \neq 0$, there is a* unique (*i.e., one and only one*) *rational number z such that $x = zy$.*

For example, making use of the Corollary to Theorem 3 of §4, we have that if $x = -\frac{1}{3}$ and $y = \frac{2}{5}$, then $z = -(\frac{1}{3} \times \frac{5}{2})$. Similarly, if $x = \frac{7}{5}$ and $y = -\frac{2}{3}$, then $z = -(\frac{7}{5} \times \frac{3}{2})$, or if $x = -\frac{7}{5}$ and $y = -\frac{2}{3}$, then $z = \frac{7}{5} \times \frac{3}{2}$. Note that, except for the negative sign, the z in all cases is obtained by invert-and-multiply.

We will reduce the proof of Theorem 1 to the special case where $x = 1$. Precisely, we first prove:

Theorem 1′ *Given any nonzero rational number y, there is a* unique (*i.e., one and only one*) *rational number z such that $zy = yz = 1$.*

Proof of Theorem 1′ First assume $y > 0$. Then y is a fraction and the existence and uniqueness of such a z is already known (see Theorem 1 of §5 in Chapter 1). If $y < 0$, then $y = -\frac{m}{n}$ for some fraction $\frac{m}{n}$. Then by Theorem 3 of the preceding section, the rational number $z = -\frac{n}{m}$ satisfies $zy = yz = 1$. If there another rational number Z so that $Zy = yZ = 1$, then

$$Z = Z \cdot 1 = Z \cdot y(-\frac{n}{m}) = (Zy)(-\frac{n}{m}) = -\frac{n}{m} = z$$

So $z = -\frac{n}{m}$ is the only number with the requisite property. The proof of Theorem 1′ is complete.

Now we give the **proof of Theorem 1** itself. With x and y given, then clearly $x = x \cdot 1 = x(y^{-1}y) = (xy^{-1})y$. So if $z = xy^{-1}$, then $x = zy$. If there is another rational number Z such that $x = Zy$, then multiplying both sides of this equation by y^{-1} yields $xy^{-1} = (Zy)y^{-1} = Z(yy^{-1}) = Z \cdot 1 = Z$, and so necessarily $Z = xy^{-1}$. Thus any such Z has to be xy^{-1}, and the proof of Theorem 1 is complete.

The number z in Theorem $1'$ so that $zy = yz = 1$ is called the **multiplicative inverse** of y (as in §5 of Chapter 1) and is denoted by $\boldsymbol{y^{-1}}$ (again as in §5 of Chapter 1). Using the concept of a multiplicative inverse, we can now prove something that is usually taken for granted.

Corollary 1 *If x, y are rational numbers and $xy = 0$ but $x \neq 0$, then $y = 0$.*

Proof Indeed, since $xy = 0$, we have $x^{-1}xy = x^{-1} \cdot 0$. The left side is $1 \cdot y$ while the right side is 0. Therefore $y = 0$. Corollary 1 is proved.

The fact implied by Corollary 1, that $xy = 0$ implies $x = 0$ or $y = 0$, is important in the solution of equations in algebra, so this fact should be carefully pointed out to students before they take up algebra.

Corollary 2 *For any nonzero rational number y, $(-y)^{-1} = -(y^{-1})$.*

Proof This can be verified separately for the case where y is positive and then for the case that y is negative y's (see Exercises 2.5 below), but it is also valuable to learn an abstract proof. Indeed, from $1 = y^{-1}y$, we get $1 = (-(y^{-1}))(-y)$ (by Theorem 3 of §4). Comparing the latter with $1 = ((-y)^{-1})(-y)$ and using the uniqueness of the multiplicative inverse of $-y$ in Theorem $1'$, we get $(-y)^{-1} = -(y^{-1})$, as claimed.

We normally omit the parentheses around y^{-1} in $-(y^{-1})$ and simply write $-y^{-1}$, and we can do this because Corollary 2 guarantees that there is no possibility of confusion.

Thus $(-\frac{2}{7})^{-1} = -\frac{7}{2}$, and $(-2\frac{1}{7})^{-1} = -\frac{7}{15}$. In general (cf. problem 1 in Exercises 2.5 below), if $\frac{m}{n}$ is any fraction,

$$\left(-\frac{m}{n}\right)^{-1} = -\frac{n}{m}$$

What does the preceding theorem really say? It says that if we have a nonzero rational number y, then *any* rational number x can be expressed as a **rational multiple** of y, in the sense that $x = zy$ for a unique *rational number* z; in fact, $z = xy^{-1}$. Thus with y fixed, every rational number x determines a unique rational number

$z = xy^{-1}$ so that $x = zy$. This existence and uniqueness of such a number z makes the following definition well-defined.

Definition *Given rational numbers x, y, with $y \neq 0$, the* **division of $\boldsymbol{x}$ by $\boldsymbol{y}$**, *in symbols,* $\dfrac{\boldsymbol{x}}{\boldsymbol{y}}$, *is the unique rational number z so that*

$$x = zy$$

The number $\frac{x}{y}$ $(= \ z)$ is also called the **quotient of $\boldsymbol{x}$ by $\boldsymbol{y}$**. Theorem 1 implies that this quotient z is given by $z = xy^{-1}$. Thus we have

$$\frac{x}{y} \ = \ xy^{-1}$$

Thus *"x divided by y" is the same as "x multiplied by the multiplicative inverse of y"*. It is in this sense that, among rational numbers, division is just multiplication, in the same way that subtraction is just addition (see §3).

We note that, as a special case, for any nonzero rational number y,

$$\frac{1}{y} \ = \ y^{-1}$$

Rational number as division of integers

We can now clear up a standard confusion in the study of rational numbers. One routinely finds in school textbooks, for instance, the equalities

$$\frac{3}{-7} \ = \ \frac{-3}{7} \ = \ -\,\frac{3}{7},$$

and they are used with nary a comment or an explanation. For example, what does something like $\frac{3}{-7}$ mean and why are the equalities true? We now supply the explanation. Because -3, -7, etc. are rational numbers, it makes sense to interpret a symbol such as $\frac{3}{-7}$ as a division of rational numbers. It then follows from the definition that

$$\frac{3}{-7} \ = \ 3 \times (-7)^{-1} \ = \ 3 \times (-\,\frac{1}{7}) \ = \ -\,\frac{3}{7}$$

where we have made use of Corollary 2 above to get $(-7)^{-1} = -\frac{1}{7}$ and Theorem 3 of §4 in the last step. In a similar fashion, we have $\frac{-3}{7} = -\ \frac{3}{7}$. More generally, the same reasoning supports the assertion that if k and ℓ are whole numbers and $\ell \neq 0$, then

$$\frac{-k}{\ell} \ = \ \frac{k}{-\ell} \ = \ - \ \frac{k}{\ell}$$

and

$$\frac{-k}{-\ell} \ = \ \frac{k}{\ell} \ .$$

We may also summarize these two formulas in the following statement:

Lemma *For any two integers a and b, with $b \neq 0$,*

$$\frac{-a}{b} = \frac{a}{-b} = - \ \frac{a}{b}.$$

This lemma will be seen to be a special case of basic facts about so-called rational quotients, to be introduced in the next sub-section, but in terms of everyday computations with rational numbers, it is well-nigh indispensable. In particular, the Lemma implies that every rational number can be written as a quotient of two integers. Because of the conceptual significance of this conclusion, we single it out in the following theorem.

Theorem 2 *Every rational number can be written as a quotient of two integers. In addition, the quotient can be chosen so that the denominator is a whole number.*

This theorem gives an alternate view of a rational number. In advanced mathematics, rational numbers are usually *defined* as quotients of integers.

Theorem 2 implies that, for instance, the rational number $-\ \frac{9}{7}$ is equal to $\frac{-9}{7}$ or $\frac{9}{-7}$, and the former is the preferred choice. This is because we know $y^{-1} = \frac{1}{y}$, so that

$$\frac{-9}{7} \ = \ -9 \times \frac{1}{7} \ = \ \frac{1}{7} \times (-9)$$

whereas

$$\frac{9}{-7} \ = \ 9 \times \frac{1}{-7} \ = \ \frac{1}{-7} \times 9$$

and it is much easier to think of $\frac{1}{7}$ of -9 than $\frac{1}{-7}$ of 9 (we have had lots of practice dividing a segment into 7 equal parts, but none dividing a segment into "-7" equal parts, whatever that means).

Rational quotients

Just as the division of fractions led to the concept of complex fractions, the division of rational numbers leads to a similar concept which, for lack of a name, will be simply referred to as **rational quotients**. We now list the analogs of the basic properties (a)–(e) of complex fractions. Let x, y, z, w, $\ldots$ be rational numbers so that they are nonzero where appropriate in the following. Then $\frac{x}{y}$ is an example of a rational quotient; x will be called its **numerator**, and y its **denominator**.

(a) Generalized cancellation law: $\boldsymbol{\frac{x}{y} = \frac{zx}{zy}}$ **for any nonzero** $\boldsymbol{z \in}$ **Q.**

(b) $\boldsymbol{\frac{x}{y} = \frac{z}{w}}$ **if and only if** $\boldsymbol{xw = yz}$**.**

(c) $\boldsymbol{\frac{x}{y} \pm \frac{z}{w} = \frac{xw \pm yz}{yw}}$.

(d) $\boldsymbol{\frac{x}{y} \times \frac{z}{w} = \frac{xz}{yw}}$**.**

(e) Distributive law $\boldsymbol{\frac{x}{y}\left(\frac{z}{w} \pm \frac{u}{v}\right) = \left(\frac{x}{y} \times \frac{z}{w}\right) \pm \left(\frac{x}{y} \times \frac{u}{v}\right)}$**.**

> *Compared with the corresponding assertions for complex fractions in §6 of Chapter 1, it will be noticed that in (b), the analog of the inequality version of the cross-multiplication algorithm is missing.* Indeed, the presence of negative numbers adds complexity to the comparison of rational numbers. This issue needs extra care and will be left as an exercise.

An immediate consequence of (a) and (d) is the **cancellation rule** among rational numbers:

(f) $\boldsymbol{\frac{x}{y} \times \frac{z}{x} = \frac{z}{y}}$.

For example, (f) justifies the cancellation $\frac{-3}{17} \times \frac{5}{-3} = \frac{5}{17}$.

As in §6 of Chapter 1, we will avoid proving (a)–(e) by the mechanical procedure of writing out each rational number as a quotient of two integers for the routine computation, but will instead make use of the *uniqueness* assertion of Theorem 1.

To prove (a), for example, let $A = \frac{x}{y}$, $B = \frac{zx}{zy}$, and we will prove that $A = B$. By the definition of division of rational numbers, we have $x = Ay$ and $zx = B(zy)$. But the first equality implies $zx = z(Ay)$ which is of course equal to $zx = A(zy)$. Now compare the latter with $zx = B(zy)$. Theorem 1 says there is only one way to express zx as a rational multiple of zy, so that we must have $A = B$.

> *We explicitly caution against incorrect reasoning* at this stage *in the passages from*
>
> $$A = \frac{x}{y} \qquad \textit{to} \qquad x = Ay,$$
>
> *and from*
>
> $$B = \frac{zx}{zy} \qquad \text{to} \qquad zx = Bzy.$$
>
> It is tempting to think that each is the result of an appropriate cancellation. For example, it would appear that by multiplying both sides of $A = \frac{x}{y}$ by y, one gets $Ay = \frac{x}{y} \times y$, so that by simplifying the right side, one arrives at $Ay = x$. However, unless we already know that (a) and (c) are true, we do not get $\frac{x}{y} \times y = x$ "by cancellation" (cf. (f) above). But we are, at this stage, still trying to prove (a), so we are in no position to do any kind of cancellation as yet. Rather, the equality $x = Ay$ is the result of the *definition* of the division of x by y. Similarly, one obtains $zx = Bzy$ from $B = \frac{zx}{zy}$ by virtue of the *definition* of dividing zx by zy.
>
> We repeat, *there is no "cancellation" in the preceding proof of* (a). But of course, once we have proved (a)–(f), we can cancel as much as we want.

To prove (d), let $A = \frac{x}{y}$, $B = \frac{z}{w}$, and $C = \frac{xz}{yw}$. We want to show $AB = C$. Again, by the definition of division, we get, respectively,

$$\begin{aligned} Ay &= x \\ Bw &= z \\ C(yw) &= xz \end{aligned}$$

Multiplying the first and second equalities together, we get $AB(yw) = xz$. Comparing the latter with the third equality, we get $AB = C$ by appealing to the uniqueness part of Theorem 1 on how to express xz as a rational multiple of yw.

The proofs of (b) and (c) are similar and will be left as an exercise, and (e) requires no proof as it is just the ordinary distributive law expressed in terms of rational quotients.

These formulas may seem unnecessarily abstract, but they have interesting, practical consequences. For example, let x, y, ... be rational numbers as before. Then

$$\left(\frac{x}{y}\right)^{-1} = \frac{y}{x}$$

This is because, by (d) and (a), $\frac{y}{x} \times \frac{x}{y} = 1$. Also, we have the **general form of invert and multiply**:

$$\frac{\frac{x}{y}}{\frac{z}{w}} = \frac{x}{y} \times \frac{w}{z}$$

This is because, by the definition of division, the left side is $\frac{x}{y}\left(\frac{z}{w}\right)^{-1}$, and because $\left(\frac{z}{w}\right)^{-1} = \frac{w}{z}$.

In school textbooks, the following kind of invert-and-multiply on rational quotients is routinely performed:

$$\frac{\frac{-3}{5}}{\frac{2.4}{-7}} = \frac{(-3)(-7)}{5 \times 2.4}$$

At the same time, students are only told to invert and multiply *ordinary fractions* (see §5 of Chapter 1). The cumulative effect of these gaps between what we expect students to know and what we actually teach them cannot help but inspire distrust. Students will second-guess everything they are taught, and "improve" on theorems they know as they go along. Under these circumstances, precise definitions and precise theorems go out the window. Such a climate is not conducive to good mathematics education.

Be sure to point out to your students that there is substantive mathematics reasoning (encoded in the Lemma, Theorem 2, and the rules (a) and (d) for rational quotients) behind the seemingly simple general invert-and-multiply rule.

Exercise 2.5

1. Give a direct proof of $(-x)^{-1} = -\left(x^{-1}\right)$ by considering the two cases separately: (i) x is a fraction and (ii) x is a negative fraction.

2. Write down an explanation you would give to an seventh grader that $-\frac{4}{5} = \frac{4}{-5}$. Expect this seventh grader to be hazy about all these symbols to begin with.

3. Explain as if to a seventh grader why $3/\frac{4}{-5} = -\frac{15}{4}$. Assume only a knowledge of the multiplication of rational numbers, and explain what division means.

4. (a) Prove that, for rational numbers x, y, z, w $(yw \neq 0)$, $\frac{x}{y} = \frac{z}{w}$ if and only if $xw = yz$. (b) Give a proof of $\frac{x}{y} + \frac{w}{z} = \frac{xz+wy}{yz}$ for rational numbers x, y, z, w $(yw \neq 0)$, *by making use of the uniqueness assertion of Theorem 1.* (See the above proofs of (a) and (c).)

5. Let x, y, z be rational numbers so that $z = \frac{x}{y}$. Explain as if to a seventh grader why (a) if x and y are both positive or both negative, z is positive, and (b) if one of x and y is positive and the other negative, then z is negative.

6. Compute and simplify: (a) $(\frac{-39}{8} \times \frac{9}{11}) + (\frac{39}{-8} \times \frac{-5}{33})$. (b) $\frac{7}{1.2} + \frac{5}{-1.8}$. (c) $-6\frac{1}{4} - \frac{27}{8}(\frac{2}{3} - \frac{8}{9})$. (d) $(-4.79) \times 0.25 - (-0.5)(1.87)$. (e) $\frac{9}{-3\frac{1}{2}} + \frac{26.7}{10.5}$.

7. (a) Find a rational number x so that $4 - \frac{5}{7}x = -3x + 2\frac{3}{4}$. (b) Show that if A, B,

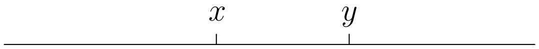
C, and D are rational numbers, and $A - C \neq 0$, then there is a rational number x so that $Ax + B = Cx + D$.

8. (a) Let x be a nonzero rational number. Explain why $\frac{x}{0}$ cannot be defined. (*Hint:* Look carefully at the definition of a division $\frac{x}{y}$ and see where the reasoning begins to break down if $y = 0$.) (b) Explain why $\frac{0}{0}$ cannot be defined. (*Caution:* This requires more care.)

6 Comparing rational numbers

The basic inequalities

Recall the definition of $\boldsymbol{x < y}$ between two rational numbers x and y (see §1): it means x is to the left of y on the number line.



We also write $\boldsymbol{y > x}$ for $x < y$. A related symbol is $\boldsymbol{x \leq y}$ (or, $y \geq x$), which means $x < y$ or $x = y$.

In this section, we will take a serious look at the comparison of rational numbers and prove several basic inequalities that are useful in school mathematics. In general, we use the symbol "<" exclusively, but you should be aware that *every one of these inequalities has an analogous statement about "≤".*

We take note of three simple properties of the inequality between numbers. The first two are the following:

If $x \le y$ and $y \le x$, then $x = y$.

If $x \le y$ and $y \le z$, then $x \le z$.

The third property deserves to be singled out because it plays a critical role in many proofs. Given any two numbers x and y, then either they are the same point, or if they are distinct, one is to the left of the other, i.e., x is to the left of y, or y is to the left of x. These three possibilities are obviously mutually exclusive. In symbols, this becomes:

> **Trichotomy law** *Given two numbers x and y, then one and only one of the three possibilities holds: $x = y$, or $x < y$, or $x > y$.*

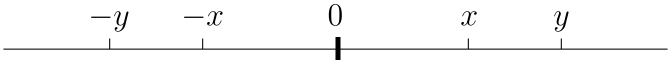
The way this law comes up in proofs is typically the following. Suppose we try to prove that two numbers x and y are equal. Sometimes it is impossible or difficult to directly prove $x = y$. But by the trichotomy law, if we can eliminate $x < y$ and $x > y$, then the fact that $x = y$ will follow.

The basic inequalities we are after are labeled (A) to (G) below. (Recall that "$\Longleftrightarrow$" stands for "is equivalent to".)

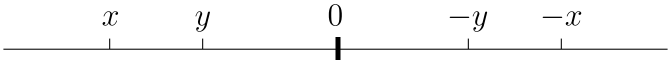
(A) For any $x, y \in \mathbf{Q}$, $x < y \iff -x > -y$.

For example, $2 < 3 \iff -3 < -2$.

If $x < 0 < y$, then $-x > 0$ while $-y < 0$ and there is nothing to prove. Therefore we need only to attend to the cases where x and y **have the same sign**, i.e., are both positive or both negative. If $0 < x < y$, then we have



On the other hand, if $x < y < 0$, then we have



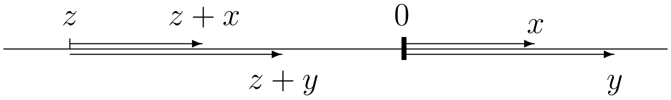
In both cases, the truth of $-x > -y$ is obvious.

(B) For any $x, y, z \in \mathbf{Q}$, $x < y \iff x + z < y + z$.

For example, given $2 < 3$, we can verify by direct computation that $2-15 < 3-15$ and $2 + \frac{7}{3} < 3 + \frac{7}{3}$.

We first prove that $x < y$ implies $x + z < y + z$ for any z. So suppose $x < y$. Because of the commutativity of addition, it suffices to prove $z + x < z + y$, or equivalently, the endpoint of the vector $\overline{z} + \overline{x}$ is to the left of the endpoint of the vector $\overline{z}+\overline{y}$. By the definition of vector addition, both vectors $\overline{z}+\overline{x}$ and $\overline{z}+\overline{y}$ are obtained by placing the starting points of $\overline{x}$ and $\overline{y}$, respectively, at the endpoint of $\overline{z}$, and the endpoints of the displaced $\overline{x}$ and $\overline{y}$, respectively, will be $z+x$ and $z+y$. Since by hypothesis, the endpoint of $\overline{x}$ is to the left of the endpoint of $\overline{y}$, the conclusion is immediate.

The following picture shows the case where $x > 0$ and $y > 0$ (and whether z is positive or negative is irrelevant):



Next we prove $x + z < y + z$ for some z implies that $x < y$. To do this, we make use of what we have just proved: adding $-z$ to both sides of $x + z < y + z$ immediately yields $x < y$. The proof of (B) is complete.

Corollary *For any $x, y, w, z \in \mathbf{Q}$, if $x < y$ and $w < z$, then $x + w < y + z$.*

The detailed proof will be left as an exercise.

(C) For any $x, y, \in \mathbf{Q}$, $x < y \iff y - x > 0$.

For example, $(-5) < (-3) \implies (-3) - (-5) > 0$ (check: $(-3) - (-5) = 2$), and conversely, $(-3) - (-5) > 0 \implies (-5) < (-3)$.

First, we prove that $x < y \implies y - x > 0$. By (B), $x < y$ implies $x + (-x) < y+(-x)$, which is equivalent to $0 < y-x$. Conversely, we prove $y-x > 0 \implies x < y$.

Again we use (B): $y - x > 0$ implies that $(y - x) + x > 0 + x$, which is equivalent to $y > x$, as desired.

It should be remarked that in higher mathematics, (C) is taken as the definition of $x < y$.

(D) For any $x, y, z \in \mathbf{Q}$, if $z > 0$, then $x < y \iff xz < yz$.

Thus, $4 < 5 \implies (\frac{23}{6})4 < (\frac{23}{6})5$ (check: the left side is $\frac{92}{6}$ and the right side is $\frac{115}{6}$). Also, $(-11) < (-9) \implies 7(-11) < 7(-9)$ (check: the left side is -77 while the right side is -63).

We first prove that, with x, y, z as given, $x < y \implies xz < yz$. We give two proofs.

First, we make use of (C). By (C), $xz < yz$ is equivalent to $(yz - xz) > 0$. Now $(yz - xz) = (y - x)z$. But we know $z > 0$ by hypothesis, and $y - x > 0$ because of the hypothesis that $x < y$ and because of (C). Since the product of two positive numbers is positive, we have $(y-x)z > 0$, which means $(yz-xz) > 0$,[36]and therefore $xz < yz$ (by (C) again), as claimed. A second proof uses Theorem 2 of §4 in Chapter 1 on fraction multiplication, which equates a product with the area of a rectangle. Given $z > 0$ and $x < y$. If $x < 0 < y$, then $xz < 0$ and $yz > 0$ and there would be nothing to prove. Therefore we need only consider the cases where x and y have the same sign (which, we recall, means they are both positive or both negative). If $x, y > 0$, then this inequality is exactly inequality (A) at the end of §4 in Chapter 1. Briefly, the proof goes as follows: x, y, and z are fractions and xz and yz are then areas of rectangles with sides of length x, z and y, z, respectively (Theorem 2 in §4 of Chapter 1). Since $x < y$, clearly the rectangle corresponding to yz contains the rectangle corresponding to xz and therefore has a greater area. Hence $yz > xz$. Next, suppose $x, y < 0$, then we get $(-x), (-y) > 0$. Moreover $x < y$ implies $(-y) < (-x)$, by (A). Thus we know from the preceding argument that $(-y)z < (-x)z$, which is equivalent to $-yz < -xz$ (Theorem 3 in §4 of this chapter), and therefore $yz > xz$, by (A) again.

[36] Note that from now on, we are secure enough about the multiplication of rational numbers that we will write $\boldsymbol{-xz}$ in place of $-(xz)$ without fear of confusion. See comment after the Corollary to Theorem 3 in §4.

Finally, we prove the converse: if for some $z > 0$, $xz < yz$, then $x < y$. We claim that $\frac{1}{z} > 0$. Indeed, since $z(\frac{1}{z}) = 1$, and $1 > 0$, we see that the product of the positive number z with $\frac{1}{z}$ is positive. Therefore $\frac{1}{z}$ has to be positive. Such being the case, then by what we have just proved, $\frac{1}{z} > 0$ and $xz < yz$ imply that $\frac{1}{z}(xz) < \frac{1}{z}(yz)$, which is the same as $x < y$. (D) is proved.

As a **Corollary**, we have: *If x, y, z, w are fractions and $x \leq y$ and $z \leq w$, then $xz \leq yw$.* The proof will be left as an exercise.

(E) For any $x, y, z \in \mathbf{Q}$, if $z < 0$, then $x < y \iff xz > yz$.[37]

To students, the fact that, when $z < 0$, the inequality $x < y$ would turn into $xz > yz$ is the most fascinating aspect about inequalities. This goes against everything they have learned up to this point, which suggests that whatever arithmetic operation they apply to an inequality, the inequality will stay unchanged. Here is a situation where an inequality gets reversed. We first illustrate with some examples. In each of the following cases, the initial inequality is multiplied by -4:

$$\begin{array}{ccc} 1 < 2 & \text{but} & -4 > -8, \\ \frac{3}{2} < \frac{15}{4} & \text{but} & -6 > -15, \\ -2 < \frac{1}{2} & \text{but} & 8 > -2, \\ -1 < -\frac{2}{3} & \text{but} & 4 > 2\frac{2}{3}. \end{array}$$

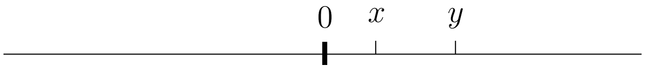
Again, we give two proofs of $x < y \implies xz > yz$ when $z < 0$. First we make use of (C). By (C), this is equivalent to proving that $xz - yz > 0$, i.e., $(x-y)z > 0$. From the hypothesis $x < y$ and using (C), we have $y - x > 0$, which implies $-(y - x) < 0$, by (A), and therefore $-y + x < 0$. In other words, $x - y < 0$. Since z is also negative, the product of the two negative numbers z and $x - y$ is positive, i.e., $(x - y)z > 0$, as desired.

For the second proof, let $z = -w$, where w is now *positive*. Since $x < y$, (D) implies that $wx < wy$. By (A), $-wx > -wy$. But Theorem 3 of §4 says $-wx = (-w)x = zx$,

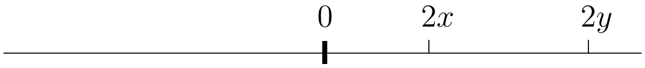
[37] In the preceding section, we warned against the tendency to assume that every skill is universally applicable. There is no better illustration of the danger of this tendency than the contrast between (D) and (E). One must begin to be sensitive to the fact that some facts are true only under restrictive hypotheses.

and $-wy = (-w)y = zy$. So $zx > zy$.

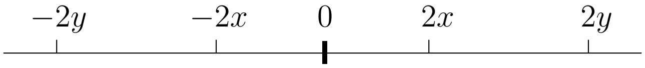
The second proof suggests a more intuitive way to understand why, if $z < 0$, then multiplying an inequality by z will reverse that inequality. Consider the special case where $0 < x < y$ and $z = -2$. So we want to understand why $(-2)y < (-2)x$. By Theorem 3 of §4, $(-2)y = -(2y)$ and $(-2)x = -(2x)$. Thus we want to see, intuitively, why $-2y < -2x$. From $0 < x < y$, we get the following picture:



Then the relative positions of $2x$ and $2y$ do not change as each of x and y is pushed further to the right of 0 by the same factor of 2. (Of course, if z were $\frac{1}{2}$, then x and y would be pushed closer to 0 by the same factor of $\frac{1}{2}$, so their relative positions would still be the same.)



If we reflect this picture across 0, we get the following:



We see that $-2y$ is now to the left of $-2x$, so that $-2y < -2x$, as claimed.

It remains to prove that if $z < 0$, then $xz > yz$ implies $x < y$. We claim that $\frac{1}{z} < 0$. This is because $z(\frac{1}{z}) = 1$ and 1 is positive. Since z is negative, $\frac{1}{z}$ has to be negative too. Thus by the first part of the proof, multiplying both sides of $xz > yz$ by $\frac{1}{z}$ would reverse the inequality, i.e., $\frac{1}{z}(xz) < \frac{1}{z}(yz)$. This is the same as $x < y$. The proof of (E) is complete.

(F) For any $x \in \mathbf{Q}$, $x > 0 \iff \frac{1}{x} > 0$.

This has essentially been proved in the course of proving (E). Observe that $x(\frac{1}{x}) = 1 > 0$. Therefore x and $\frac{1}{x}$ are either both positive, or both negative, as claimed.

The next item is an immediate consequence of (D)–(F).

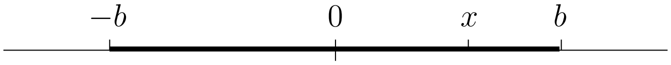
(G) For any $x, y, z \in \mathbf{Q}$, let $x < y$. If $z > 0$, then $\frac{x}{z} < \frac{y}{z}$; if $z < 0$, then $\frac{x}{z} > \frac{y}{z}$.

Absolute value

Next, we turn to the concept of the absolute value of a number, which is intrinsically tied to any discussion of inequalities. The **absolute value $|x|$** of a number x is by definition the **distance** from x to 0 (i.e., the length of the segment $[x, 0]$ or $[0, x]$, depending on whether x is negative or positive, respectively). In particular, $|x| \geq 0$ no matter what x may be. The most pleasant property of the absolute value is that, for all numbers x, y,

$$|x|\,|y| = |xy|$$

This can be proved by a case-by-case examination of the four cases where x and y take turns being positive and negative. The reasoning is routine. On the other hand, inequalities involving absolute value present difficulties to students, so it is absolutely essential that we come to grips with such inequalities. If b is a positive number, then the set of all numbers x so that $|x| < b$ consists of all the points x of distance less than b from 0, indicated by the thickened segment below (excluding the endpoints):



It follows that the inequality $|x| < b$ for a point x is equivalent to the fact that x satisfies both $-b < x$ and $x < b$. It is standard practice in mathematics to combine these two inequalities into a composite statement in the form of a **double inequality** $\boldsymbol{-b < x < b}$. In this notation, we can summarize what we have proved neatly, as follows:

$$|x| < b \textit{ is equivalent to } -b < x < b.$$

In the usual notation for intervals on the number line, this becomes:

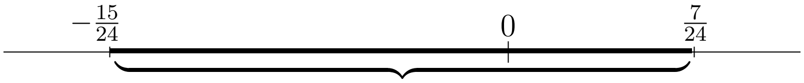
$$|x| < b \quad \textit{is equivalent to} \quad x \in (-b, b).$$

> Recall: the set of all the points x satisfying $c < x < d$, where c and d are two fixed numbers so that $c < d$, is denoted by $\boldsymbol{(c, d)}$, called an **open interval**. The segments we have been using thus far are sets of the form (c, d) together with the endpoints c and d; these are denoted by $\boldsymbol{[c, d]}$, called a **closed interval** , i.e., $[c, d]$ is the set of all points x so that $c \leq x \leq d$.

The fact that *the single inequality* $|x| < b$ *involving absolute value is equivalent to a double inequality* $-b < x < b$ is a very useful fact in the elementary considerations involving absolute value. In the following, we sometimes refer to $-b < x < b$ as the **associated double inequality** of $|x| < b$. The following example illustrates the way the conversion of an absolute value inequality into its associated double-inequality can be put to use.

Example 1 Determine all the numbers x so that $|6x + 1| + 2\frac{1}{4} < 5$, and show them on the number line.

The inequality $|6x + 1| + 2\frac{1}{4} < 5$ is equivalent to $|6x + 1| < 5 - 2\frac{1}{4}$, (by (B) above), which is just $|6x + 1| < 2\frac{3}{4}$, which in turn is equivalent to the double inequality $-2\frac{3}{4} < 6x + 1 < 2\frac{3}{4}$. The left inequality is equivalent to $-2\frac{3}{4} - 1 < 6x$ (by (B) again), i.e., $-\frac{15}{4} < 6x$. Now we multiply both sides of this inequality by $\frac{1}{6}$ and use (D) to conclude that it is equivalent to $-\frac{15}{24} < x$. By exactly the same reasoning, the right inequality $6x + 1 < 2\frac{3}{4}$ is equivalent to $x < \frac{7}{24}$. Putting all this together, we have that the inequality $|6x + 1| + 2\frac{1}{4} < 5$ is equivalent to the double inequality $-\frac{15}{24} < x < \frac{7}{24}$. The set of all x satisfying this double inequality is indicated by the thickened segment in the picture.

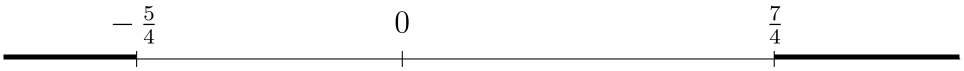


Example 2 Determine all the numbers x so that $|2x - \frac{1}{2}| \geq 3$, and show them on the number line.

There are two cases to consider: *Case I*: $2x - \frac{1}{2} \geq 0$, and *Case II*: $2x - \frac{1}{2} < 0$. First we look at Case I. Then the inequality becomes $2x - \frac{1}{2} \geq 3$, $\Longleftrightarrow$ $2x \geq \frac{7}{2}$ (by (B)), $\Longleftrightarrow$ $x \geq \frac{7}{4}$ (by (D)). Therefore, in case $2x - \frac{1}{2} \geq 0$, $|2x - \frac{1}{2}| \geq 3$, if and only if $x \geq \frac{7}{4}$.

Next, suppose $2x - \frac{1}{2} < 0$. Then $|2x - \frac{1}{2}| = \frac{1}{2} - 2x$, so that $|2x - \frac{1}{2}| \geq 3$ $\Longleftrightarrow$ $\frac{1}{2} - 2x \geq 3$, $\Longleftrightarrow$ $-2x \geq \frac{5}{2}$, $\Longleftrightarrow$ $x \leq -\frac{5}{4}$ (by (E)). Therefore in case $2x - \frac{1}{2} < 0$, $|2x - \frac{1}{2}| \geq 3$, if and only if $x \leq -\frac{5}{4}$.

Together, we see that $|2x - \frac{1}{2}| \geq 3$ if and only if either $x \geq \frac{7}{4}$ or $x \leq -\frac{5}{4}$, corresponding to either $2x - \frac{1}{2} \geq 0$, or $2x - \frac{1}{2} < 0$. On the number line, the numbers satisfying the inequality are exactly those numbers lying in one of the two thickened semi-infinite segments.



Having introduced the concept of absolute value, we now face the question, asked by most teachers (not to mention innumerable students), of why we bother with this concept. As one educator noted, in schools, absolute value is often taught as a topic disconnected from anything else in the curriculum; it is barely touched on, and is not understood except as a kind of rote procedure ("take off the minus sign if there is one"). Teachers feel handicapped by being made to teach something for which they don't see any relevance.

It is not possible in an elementary text to give a wholly satisfactory answer to the question of why absolute value should be taught. The importance of absolute value emerges mostly in the more advanced portion of mathematics or the sciences, such as when we come face-to-face with the concept of limit and the unavoidable inequalities or when making estimates becomes a necessity rather than empty rhetoric. In these notes, however, we have to be content with giving only a rough idea of its significance.

There are situations where we want only the absolute value ("magnitude") of a number, but do not care much whether the number is positive or negative. For example, suppose you try to estimate the sum of two 3-digit whole numbers, $369 + 177$, by rounding to the nearest hundred. The sum is of course 546, but the estimated sum would be $400 + 200 = 600$. The measurement of the accuracy of such an estimate is the so-called **absolute error** of the estimation, which is by definition the *absolute value* of the difference between the true value and the estimated value, i.e.,

$$\textbf{absolute error} = |\textit{true value} - \textit{estimated value}|$$

In this case, it is $|546 - 600| = 54$. Now, if we do the same to the sum $234 + 420$, then the absolute error of the estimated value of 600 ($= 200 + 400$) is still 54, because $|654 - 600| = 54$. These two estimates differ in that the former under-estimates by an amount of 54, whereas the latter over-estimates by the same amount. However, as a preliminary indication of the accuracy of these estimates, it can be said that they *both miss the mark by 54* and it doesn't matter whether they are over or under by this amount. Thus it is the absolute value of this difference, rather than the difference itself, that is of primary interest. The absolute value in this instance provides exactly the right tool to express the *error* of such estimations.

Two useful inequalities

Another way to see why absolute value is essential is to see how it is used. We illustrate with two standard inequalities. Here is the first.

Theorem 1 *For any numbers x and y,*

$$2|xy| \;\leq\; x^2 + y^2$$

By FASM, it suffices to prove this for rational values of x and y, and we will tacitly assume $x, y \in \mathbf{Q}$ in the discussion following. Before giving the proof, there are at least two things one should do. The obvious one is to become convinced, psychologically, that this inequality is true. If you have no conviction, then there is no hope of arriving at a proof. For inequalities, conviction comes readily enough by trial and error: let x and y be replaced by concrete numbers to see if the inequality still holds. For example, if $x = 7$ and $y = 11$, the inequality suggests $154 \leq 170$. If $x = 8$ and $y = 10$, the inequality gives $160 \leq 164$. And so on. Now if $x = y = 9$, the inequality becomes $162 \leq 162$, and the same thing happens if $x = y = 8$: $128 \leq 128$, etc. In fact, once this observation is made, you begin to notice that if $x = y$, then both sides of the inequality become $2x^2$ (there is no need for the absolute value in this case because $|x|^2 = |x^2| = x^2$.) Now you get to see why "$\leq$" is used rather than "$<$". We will come back to the case of equality presently.

Another thing one should do before attempting any proof is to at least understand why the absolute value $|xy|$ of xy is used at all. Clearly, this theorem is of no interest if one of x and y is 0, as it merely says in that case that $0 \leq x^2 + y^2$. So we may as well assume both x and y to be nonzero. Such being the case, we make use of (G) above to rewrite the theorem as

$$\frac{2|xy|}{x^2 + y^2} \;\leq\; 1$$

for all x and y. Since $x^2 + y^2 \geq 0$, we have $\frac{1}{x^2+y^2} \;\geq 0$ on account of (F). It follows that $\left|\frac{1}{x^2+y^2}\right| = \frac{1}{x^2+y^2}$. Therefore, using $|AB| = |A|\,|B|$ for all numbers A and B, we get

$$\left|\frac{2xy}{x^2 + y^2}\right| = |2xy| \cdot \left|\frac{1}{x^2 + y^2}\right| = 2|xy| \cdot \frac{1}{x^2 + y^2} \;=\; \frac{2|xy|}{x^2 + y^2}$$

Thus the theorem is equivalent to

$$\left|\frac{2xy}{x^2 + y^2}\right| \;\leq\; 1$$

We know from a previous remark that this inequality is equivalent to the double inequality

$$-1 \;\leq\; \frac{2xy}{x^2 + y^2} \;\leq\; 1$$

In this form, the theorem asserts that the number $\frac{2xy}{x^2+y^2}$ is trapped inside the segment $[-1, 1]$ between -1 and 1 for all x and y.

Without the absolute value sign, the theorem merely says that

$$\frac{2xy}{x^2 + y^2} \;\leq\; 1$$

This inequality does not preclude the possibility that $\frac{2xy}{x^2+y^2} \;= -100$. With the absolute value sign in place, however, we know that $\frac{2xy}{x^2+y^2}$ cannot be to the left of -1 on the number line and, in particular, cannot be equal to -100. We see clearly that, in this case, the presence of absolute value in the inequality is an inequality that carries much more information than the one without absolute value.

The preceding discussion gives the barest glimpse into what happens in advanced mathematics, where very often we want to control the absolute value of a given number, in much the same way that we want to control the absolute value of xy in Theorem 1. Typically, this control is obtained only after stringing together a sequence of inequalities involving absolute values. If we do not explicitly make use of absolute value at each stage, then we will be forced to deal instead with *two* inequalities each time (i.e., those given by the associated double inequalities). As the number of such inequalities involving absolute value increases, the number of ordinary inequalities we need to look at becomes unmanageable. The use of absolute value is thus a necessity.

It remains to give the simple **proof of Theorem 1**. We prove it in its original formulation:

$$2|xy| \;\leq\; x^2 + y^2$$

Let $u = |x|$ and $v = |y|$, then as we have seen, $2|xy| = 2|x|\,|y| = 2uv$. Now we make the simple observation that for all numbers t, $t^2 = |t|^2$; this is clear when we first consider the case $t \geq 0$ and then the case $t < 0$. Therefore, we have $x^2 = |x|^2 = |x|\,|x| = uu = u^2$. Similarly, $y^2 = v^2$. Thus the inequality becomes

$$2uv \;\leq\; u^2 + v^2$$

which is equivalent to $0 \leq u^2 - 2uv + v^2$, by (B) above. In other words, the theorem is equivalent to

$$u^2 - 2uv + v^2 \;\geq\; 0$$

for any numbers u and v. This is however obvious because $u^2 - 2uv + v^2 = (u - v)^2$ and $(u - v)^2 \geq 0$. The proof is complete.

At this point, we pick up the thread of the discussion right after the statement of Theorem 1 about when *equality* actually takes place in the inequality, i.e., if two numbers x and y satisfy

$$2|xy| \;=\; x^2 + y^2,$$

what can we say about them? If we let $u = |x|$ and $v = |y|$ as in the preceding proof, then we have $2uv = u^2 + v^2$, which becomes $u^2 - 2uv + v^2 = 0$. Thus $(u - v)^2 = 0$.

But the square of any number t is ≥ 0 (see remark after the Corollary to Theorem 3 in §4), and is 0 if and only if $t = 0$, so $u - v = 0$, i.e., $u = v$, or $|x| = |y|$. Conversely, if $|x| = |y|$, then both sides of the inequality in Theorem 1 are equal to $2|x|^2$. Therefore we see that the inequality in Theorem 1 is an equality if and only if $|x| = |y|$. We may therefore restate Theorem 1 in a more refined version, as follows.

Theorem 1* *For any numbers x and y,*

$$2|xy| \;\leq\; x^2 + y^2$$

The inequality is an equality, i.e., $2|xy| = x^2 + y^2$, if and only if $|x| = |y|$.

You may consider this excursion into considering the extreme case of equality a mildly entertaining exercise, but no more. What actually happens is that, most of the time, when a (weak) inequality becomes an equality, something dramatic happens. In the case of Theorem 1, this is no exception. See Chapter 6.

We conclude with what is probably the most basic inequality involving absolute value in elementary mathematics.

Theorem 2 (Triangle Inequality) For any numbers x and y,

$$|x + y| \;\leq\; |x| + |y|$$

In this case, the critical role played by absolute value is all too obvious as the inequality would be meaningless without the absolute value symbols. Incidentally, there will be a "real" triangle inequality in §1 of chapter 5 (see (D4) of that section). The reason Theorem 2 is called the Triangle Inequality has to do with fact that if x and y are vectors and the absolute value symbol is interpreted as length, then Theorem 2 would be a statement about triangles.

Proof If one of x and y is 0, then there is nothing to prove. We assume therefore that both x and y are nonzero. The most elementary proof is one using case-by-case examination of the inequality. There are two cases to consider: (i) x and y have the same sign (recall: this means both are positive or both are negative), and (*ii*) x and y have **opposite signs** (i.e., one is positive and the other is negative). Each case then splits into sub-cases. Such a proof is boring (compare the proof of Theorem 4 in §2), but it does have two things in its favor. Such a proof is instructive if you want to get some down-to-earth feelings about absolute values, and furthermore, it is the easiest way to analyze the situation when equality takes place. (See below for the latter.)

We give a different proof, one that makes use of the fact that the inequality $|x| \leq b$ is equivalent to the double-inequality $-b \leq x \leq b$. This is a standard proof, but is also one from which one can learn something about absolute values. Therefore instead of proving $|x+y| \leq |x| + |y|$, we prove the double inequality

$$-|x| - |y| \ \leq \ x + y \ \leq \ |x| + |y|$$

There is no question that $-|x| \leq x \leq |x|$ and $-|y| \leq y \leq |y|$. From $-|x| \leq x$ and $-|y| \leq y$, we use the corollary of (B) above to conclude that $-|x| - |y| \leq x + y$. Similarly, we use $x \leq |x|$ and $y \leq |y|$ and the corollary of (B) to conclude that $x + y \leq |x| + |y|$. Thus we have proved both inequalities in the double inequality.

As in the case of Theorem 1, we ask when does equality take place for the triangle inequality? If both x and y are positive or if both are negative, then clearly we have equality. If one of x and y is 0, then again we have equality. Suppose exactly one of x and y is positive and the other negative. Because the inequality doesn't change if x and y are interchanged, we may assume $x > 0$ and $y < 0$ without loss of generality. If $x + y > 0$, then

$$|x+y| \ = \ x + y \ < \ x + 0 \ < \ x \ = \ |x| \ < \ |x| + |y|$$

and equality is impossible in the triangle inequality. If $x + y < 0$, then

$$|x+y| \ = \ -(x+y) \ = \ -y - x \ < \ -y \ = \ |y| \ < \ |x| + |y|$$

and again equality is impossible. Thus we have proved:

The triangle inequality is an equality if and only if one of the numbers is 0, or the numbers have the same sign.

Exercises 2.6

Here as later, a declarative statement means simply that you have to supply a proof.

1. (a) If x, y, z, w are rational numbers and $x \leq y$ and $w \leq z$, then (a) $x+w \leq y+z$. (This is the Corollary of (B).) (b) If, in addition, all four numbers ≥ 0, then $xw \leq yz$.

2. (a) Let $x, y, z, w \in \mathbf{Q}$, and let $y, w > 0$. Then prove that $\frac{x}{y} < \frac{z}{w} \iff xw < yz$. (b) Give examples to show that both implications "$\frac{x}{y} < \frac{z}{w} \implies xw < yz$" and "$xw < yz \implies \frac{x}{y} < \frac{z}{w}$" are false without the assumption that $y, w > 0$. (c) Are the following numbers

$$\frac{32.5}{-3} \quad \text{and} \quad \frac{-30\frac{2}{3}}{2\frac{4}{5}}$$

equal? If so, prove. If not, which is bigger?

3. Which is greater? (a) $(-1.7)(9)$ or $-22 + 6\frac{2}{3}$. (b) $\frac{-2}{5}$ or $(-5)\frac{1.1}{12.5}$. (c) $\frac{-2}{3}/\frac{4}{7}$ or $(\frac{14}{3})(\frac{-2}{8.5})$.

4. (a) Determine all the numbers x which satisfy $|x-1| - 5 < \frac{2}{3}$ and show them on the number line. (b) Do the same for $11 - |3+2x| > 2.5$. (c) Do the same for $|2x - \frac{3}{5}| \geq \frac{1}{5}$. (d) Do the same for $3 - |2x-5| \geq 4.2$.

5. For any two rational numbers p and q, the length of the segment between p and q is exactly $|p-q|$.

6. Let x and y be rational numbers. (a) How does $|x| - |y|$ compare with $|x-y|$? Why? (b) How does $|\,|x| - |y|\,|$ compare with $|x-y|$? Why?

7. If x and y are rational numbers, and $y \neq 0$, then

$$\left|\frac{x}{y}\right| = \frac{|x|}{|y|}$$

In other words, the absolute value of x/y is equal to the quotient of $|x|$ by $|y|$.

8. If x and y are positive rational numbers, then (a) $x^2 = y^2$ if and only if $x = y$, and (b) $x^2 < y^2$ if and only if $x < y$.

9. If x is a rational number, is it true that $x < 1$ implies $\frac{1}{x} > 1$? If so, prove. If not, formulate a true statement, and prove it.

10. If x, y are numbers so that $0 < x < y$, and n is a positive integer, how does x^n compare with y^n ? Why?

11. If $x > 1$, then $x^n > 1$ for any positive integer n. Also if $-1 < x < 1$, then $-1 < x^n < 1$ for any positive integer n.

12. Let x be a rational number. (a) If $x > 1$, then $x^m > x^n$ for whole numbers $m > n$. (b) If $0 < x < 1$, then $x^m < x^n$ for whole numbers $m > n$.

13. For any two positive numbers s and t, $s + t \geq 2\sqrt{st}$. Furthermore, equality holds if and only if $s = t$.

14. Show that for all numbers x, y, and $c \neq 0$,

$$|x+y|^2 \;\leq\; (1 + \frac{1}{c^2})\,|x|^2 + (1 + c^2)\,|y|^2$$

15. Can you see why if x and y are any two rational numbers (in particular, they could be negative), then $\frac{1}{9}x^2 - \frac{1}{12}xy + \frac{1}{64}y^2 \;\geq\; 0$?

Chapter 3: The Euclidean Algorithm

This chapter gives two principal applications of the Euclidean Algorithm, which is a finite procedure for finding the greatest common divisor of two given whole numbers.

1 The reduced form of a fraction

A fraction $\frac{m}{n}$ is said to be a **reduced form** of a given fraction $\frac{k}{\ell}$ if $\frac{m}{n} = \frac{k}{\ell}$ and if no whole number other than 1 divides both the numerator m and the denominator n. In general, a fraction with the property that no whole number other than 1 divides both the numerator and the denominator is said to be **in lowest terms**, or **reduced**. A fact taken for granted in elementary school is that any fraction has a reduced form, and that there is only one. When classroom instruction focuses entirely on fractions with single-digit numerator and denominator, the reduced form of a fraction can be obtained by visual inspection. For fractions with larger numerators and denominators, deciding whether a fraction is in reduced form is often not so obvious. For example, is the fraction

$$\frac{1147}{899}$$

reduced? (It is not. See Exercises 3.1.)

The purpose of this section is to clarify this situation once and for all by proving the following theorem. The statement requires that we introduce the term **algorithm**, which is *an explicit finite procedure that leads to a desired outcome.*

Theorem 1 *Every fraction has a unique reduced form. Furthermore, this reduced form can be obtained by an algorithm.*

The proof of the theorem involves some number-theoretic considerations about whole numbers. We start at the beginning.

We say a nonzero integer d is a **divisor** or a **factor** of an integer a, or ***d* divides *a***, if $a = cd$ for some integer c. We also call $a = cd$ a **factorization** of a. Another way to say d divides a is to say that the rational number $\frac{a}{d}$ is an integer. We write $\boldsymbol{d|a}$ when this happens, and we also say a is an **(integral) multiple** of d. If d does not divide a, we write $\boldsymbol{d \not| a}$.

Observe that (i) if $k|\ell$ and $\ell|m$, then $k|m$, and (ii) every nonzero integer divides 0. The simple proofs are left as exercises.

In the following discussion, *most of the time all the integers involved are whole numbers*, i.e., integers which are positive or 0. However, there are one or two places which would become very awkward if we restrict ourselves only to whole numbers (cf. the proof of the Key Lemma below). For this reason, we bring in integers from the beginning. When we need to focus on whole numbers, we will be explicit about it, e.g., the concept of a *prime* immediately following.

A whole number a which is greater than 1 has at least two whole number divisors, 1 and a itself. A **proper divisor** d of a whole number a is a whole number divisor of a so that $1 < d < a$. Note that if $a = cd$ for whole numbers c and d so that both c and d are > 1, then both c and d are proper divisors of a. A whole number > 1 without proper divisors is called a **prime**, or **prime number**. A whole number which is > 1 and is not a prime is called a **composite**. Note that by definition, 1 is neither prime nor composite. Checking whether a whole number is a prime, while difficult in general, is easier than appears at first sight, because we have the following lemma.

Lemma *Given a whole number $n > 1$. If no prime number p satisfying $2 \leq p \leq \sqrt{n}$ is a divisor of n, then n is a prime.*

> For a positive number x, its **positive square root $\sqrt{x}$** is the positive number so that its square is equal to x, i.e., $(\sqrt{x})^2 = x$. It can be proved that any positive real number has a positive square root and there is no danger of circular reasoning if we make use of this fact here.
>
> Observe that if a, b are positive numbers, then $a < b$ is equivalent to $\sqrt{a} < \sqrt{b}$ (problem 8(b) in Exercises 2.6).

For example, to check whether 233 is a prime, it suffices to check whether any of the primes ≤ 16 divides 233 (because $16^2 = 256 > 233$). The primes in question are 2, 3, 5, 7, 11, 13. Since none of them divides 233, we conclude that 233 is a prime.

The Lemma can be rephrased equivalently as one about composite numbers: *If $n \in \mathbf{N}$ is composite, then it has a prime divisor p in the range $2 \leq p \leq \sqrt{n}$.* We will prove this version of the Lemma.

Proof Suppose n is composite. We first show that n has a proper divisor ℓ in the range $2 \leq \ell \leq \sqrt{n}$. If not, then every proper divisor of n exceeds $\sqrt{n}$. So let ℓ be a proper divisor of n, and we write $n = m\ell$ for some whole number m. Then m is itself a divisor of n and therefore $m > \sqrt{n}$ as well. We then have (cf. problem 1(b) of Exercises 2.6):

$$n = m\ell > \sqrt{n}\sqrt{n} = n$$

and therefore we have the absurd situation of $n > n$. Thus n must have a proper divisor ℓ so that $2 \leq \ell \leq \sqrt{n}$.

If ℓ happens to be a prime, we are done. If not, then among all the distinct proper divisors of ℓ, let p be the smallest. Then p must be a prime because otherwise p has a proper divisor q, and since $q|p$ and $p|\ell$, we have $q|\ell$. Since $q < p$, p is not the smallest proper divisor of ℓ, which is a contradiction. So p is a prime, and $p < \ell \leq \sqrt{n}$, as claimed. The Lemma is proved.

Given two whole numbers a and d, not both equal to 0, there is at least one whole number that divides both a and d, namely, the number 1. A whole number c is said to be the **GCD** (**greatest common divisor**) of whole numbers a and d if, among all the whole numbers which divide both a and d, c is the greatest. Notation: $\mathbf{GCD}(\boldsymbol{a}, \boldsymbol{d})$. Two whole numbers a and d (again not both equal to 0) are said to be **relatively prime** if $\mathrm{GCD}(a, d) = 1$.

We can approach GCD from another angle. Given a whole number n, let $\mathcal{D}(n)$ be the collection of all the divisors of n. Then if a and d are both whole numbers and not both equal to 0,

the collection of all common divisors of a and d $= \mathcal{D}(a) \cap \mathcal{D}(d)$

The following is therefore an equivalent formulation of the concept of GCD:

$\mathrm{GCD}(a, d) = \max\{\mathcal{D}(a) \cap \mathcal{D}(d)\}$

where **max** indicates the largest number in the set. In this formulation, it is clear that the GCD of any two whole numbers a and d always exists: after all, the set

$\mathcal{D}(a) \cap \mathcal{D}(d)\}$ is a finite set and all we have to do is pick the largest number in it. In this notation, we also see that a and d being relatively prime is equivalent to

$$\mathcal{D}(a) \cap \mathcal{D}(d) = 1$$

At this point, we can explain why we are interested in the GCD of two whole numbers. Let a fraction $\frac{m}{n}$ be given. Let k be the GCD of m and n, and let $m = km'$ and $n = kn'$ for some whole numbers m' and n', We claim that $\frac{m'}{n'}$ is the reduced form of $\frac{m}{n}$. The equality $\frac{m'}{n'} = \frac{m}{n}$ is a consequence of the Theorem on Equivalent Fractions. As to why $\frac{m'}{n'}$ is reduced, suppose it is not. Then m' and n' have a common divisor $\ell > 1$, let us say $m' = \ell a$ and $n' = \ell b$ for some whole numbers a and b. Then

$$m = km' = k\ell a, \quad \text{and} \quad n = kn' = k\ell b$$

It follows that $k\ell$ is a common divisor of m and n. Since $\ell > 1$, $k\ell > k$, and this contradicts the fact that k is the greatest of the common divisors. Our claim of the existence of a reduced form for $\frac{m}{n}$ is now proved.

From a practical point of view, what we want is more than a theoretical assurance that there is a GCD; we want an explicit procedure that unfailingly yields the GCD of two given whole numbers. This is what the Euclidean algorithm accomplishes.

The proof of the uniqueness of the reduced form turns out to be more subtle. The key to this proof is the following lemma which is also the key to many other basic facts. What is interesting is that this lemma is also a natural consequence of the Euclidean algorithm.

Key Lemma *Suppose ℓ, m, n are nonzero whole numbers, and $\ell | mn$. If ℓ and m are relatively prime, then $\ell | n$.*

One can appreciate this Key Lemma better if one notices that a whole number ℓ can divide a product without dividing either factor. Thus, $63|(18 \times 245)$, but $63 \not| 18$ and $63 \not| 245$. What the Key Lemma says is that if ℓ is relatively prime to one of the factors, then ℓ dividing the product would imply that ℓ divides the other factor. (It goes without saying that 63 is relatively prime to neither 18 nor 245, so there is no contradiction to the Key Lemma.)

The proof of the Key Lemma requires some preparation concerning the greatest common divisor of two numbers, which in turn requires that we review the well-known procedure of **division with remainder**: given whole numbers a and d, then the division with remainder of a by d is given by

$$a = qd + r \quad \text{where } q, r \in \mathbf{N}, \text{ and } 0 \leq r < d$$

The whole number r is the **remainder**. (In abstract algebra, this is of course the *division algorithm* for integers, but in school mathematics, one cannot afford to use this terminology because it causes confusion with the *long division algorithm.*) The main observation here is that, given $a = qd + r$,

$$\text{GCD}(a, d) = \text{GCD}(d, r)$$

The way we prove this is to prove something slightly more general, namely, the equality of the following two *sets*:

$$\mathcal{D}(a) \cap \mathcal{D}(d) = \mathcal{D}(d) \cap \mathcal{D}(r)$$

By the characterization of the GCD of two numbers in terms of the divisors of each number, this implies the above observation about GCD.

Before giving the proof of this equality of the sets of common divisors, we should make two remarks. First, this equality of sets means precisely that the following two inclusions hold (recall: "$\subset$" means "is contained in", and "$A \cap B$" means "the intersection of the sets A and B"):

$$\mathcal{D}(a) \cap \mathcal{D}(d) \subset \mathcal{D}(d) \cap \mathcal{D}(r)$$

and

$$\mathcal{D}(d) \cap \mathcal{D}(r) \subset \mathcal{D}(a) \cap \mathcal{D}(d)$$

A second remark is that suppose A, B, C are whole numbers and $A = B + C$. If a whole number n divides any two of A, B, C, then n divides all three. The proof is straightforward (see Exercise 3.1). This fact will be used several times in the proof of the inclusions.

Let us prove the first inclusion relationship:

$$\mathcal{D}(a) \cap \mathcal{D}(d) \subset \mathcal{D}(d) \cap \mathcal{D}(r)$$

Suppose a whole number n belongs to the left side, then we must prove that it belongs to the right side. In other words, if n divides both a and d, then it divides both d and r. Therefore we must prove that if n divides both a and d, then it divides the r in

$$a = qd + r$$

However, this means n divides both a and qd in this equation, so it divides the third number r, by the second remark above.

The proof of the reverse inclusion is entirely similar. It follows that if

$$a = qd + r,$$

then

$$\text{GCD}(a, d) = \text{GCD}(d, r)$$

There are many reasons why this fact is interesting. The most obvious reason can be seen from a simple example. Suppose we want to get the GCD of 469 and 154. We have

$$469 = (3 \times 154) \ + \ 7$$

Therefore $\text{GCD}(469, 154) = \text{GCD}(154, 7)$, and since obviously $\text{GCD}(154, 7) = 7$, it follows immediately that $\text{GCD}(469, 154) = 7$. Thus one application of division with remainder suffices to yield the GCD of 469 and 154.

Is this an accident? Not entirely, because the determination of the GCD of two numbers a and d is quite easy if at least one of a and d is sufficiently small. Witness the fact that $\text{GCD}(154, 7) = 7$. In general, if d is sufficiently small, then the set $\mathcal{D}(d)$ can be determined by visual inspection one way or another, and therewith also the set $\mathcal{D}(a) \cap \mathcal{D}(d)$. Thus $\text{GCD}(a, d)$, being $\max\{\mathcal{D}(a) \cap \mathcal{D}(d)\}$, can be determined. The virtue of having the equality $\text{GCD}(469, 154) = \text{GCD}(154, 7)$ is, therefore, that instead of having to deal with two fairly large numbers 469 and 154, we are reduced to dealing with 154 and 7, *and 7 is sufficiently small* (no matter how one defines "sufficiently small"). As we have seen, this leads immediately to the determination of $\text{GCD}(469, 154)$.

In one sense, though, this example is an accident: in this case, *one* application of division with remainder suffices to determine $\text{GCD}(469, 154)$. In general, one application is not enough. For example, suppose we try to find $\text{GCD}(3008, 1344)$. From

$$3008 = (2 \times 1344) \ + \ 320,$$

we get $\text{GCD}(3008, 1344) = \text{GCD}(1344, 320)$. This time, it is not obvious what $\text{GCD}(1344, 320)$ is, because for one thing, 320 has too many divisors so that $\mathcal{D}(320)$ becomes unwieldy. *But we can again apply division with remainder to* 1344 *and* 320 to get a yet smaller number:

$$1344 = (4 \times 320) \ + \ 64$$

and we get $\text{GCD}(1344, 320) = \text{GCD}(320, 64)$. But now $\mathcal{D}(64)$ is seen to consist of powers of 2 up to the 6th power, and it turns out that 64 itself divides 320, and therefore $\text{GCD}(320, 64) = 64$. Hence

$$\text{GCD}(3008, 1344) = \text{GCD}(1344, 320) = \text{GCD}(320, 64) = 64$$

We can further illustrate this process with a slightly more complicated example: let us find the GCD of 10049 and 1190. From

$$10049 = (8 \times 1190) \ + \ 529$$

we get $\text{GCD}(10049, 1190) = \text{GCD}(1190, 529)$. Since it is not obvious what $\text{GCD}(1190, 529)$ is, we apply division with remainder to 1190 and 529:

$$1190 = (2 \times 529) \ + \ 132$$

Now $\text{GCD}(10049, 1190) = \text{GCD}(529, 132)$, because $\text{GCD}(1190, 529) = \text{GCD}(529, 132)$. However, it is still not immediately obvious what $\text{GCD}(529, 132)$ is, so we go one step further:

$$529 = (4 \times 132) \ + \ 1$$

This time, $\text{GCD}(132, 1) = 1$, so $\text{GCD}(10049, 1190) = \text{GCD}(132, 1) = 1$. Incidentally, we have exhibited a nontrivial example of a pair of relatively prime integers: 10049 and 1190.

It is not difficult to explain this method of determining the GCD of two whole numbers in general. The idea is that the application of division with remainder to two whole numbers reduces the problem of finding the GCD of these numbers to the determination of the GCD of a second pair of numbers which are correspondingly smaller than the original pair. Thus, given the whole number pair a and d, with $a > d$, division with remainder yields

$$a = qd \ + r, \qquad \text{where } 0 \le r < d$$

The equality

$$\mathrm{GCD}(a,d) = \mathrm{GCD}(d,r)$$

replaces the determination of the GCD of a and d by the determination of the GCD of d and r, and the advantage is that $a > d$ and $d > r$. This process can be repeated by doing division with remainder on the pair d and r, getting

$$d = q_1 r \; + r_1, \qquad \text{where } 0 \leq r_1 < r$$

Now we get $\mathrm{GCD}(a,d) = \mathrm{GCD}(d,r) = \mathrm{GCD}(r,r_1)$, and $a > d > r > r_1$. In this way, we introduce smaller and smaller numbers at each step into the problem, until we inevitably reach either 1 or 0. At that point, if not earlier, we would be done because, in either case,

$$\mathrm{GCD}(n,1) = 1 \quad \text{and} \quad \mathrm{GCD}(n,0) = n$$

for any whole number n.

It is now clear that iterations of division with remainder lead to the determination of the GCD of any two whole numbers in a finite number of steps. What is even more remarkable is the fact that there is more information to be extracted from this process. First consider the simplest case of the GCD of 469 and 154. We had

$$469 = (3 \times 154) \; + \; 7$$

This equation not only shows that $\mathrm{GCD}(469, 154) = 7$, but also exhibits the GCD, which is 7, as an **integral linear combination** of the two original numbers 469 and 154, in the sense that 7 is an integer multiple of 469 plus an integer multiple of 154, namely,

$$7 = \{1 \times 469\} \; + \; \{(-3) \times 154\}\,.$$

If we consider the fact that the GCD of 469 and 154 is defined in terms of 469 and 154 using the concept of multiplication, the expression of this GCD as the *sum* of multiples of these two numbers must come as a surprise, to say the least.

Let us represent $\mathrm{GCD}(3008, 1344) = 64$ as an integral linear combination of 3008 and 1344. We first list the steps of the division with remainder:

$$\begin{aligned} 3008 &= (2 \times 1344) \; + \; 320 \\ 1344 &= (4 \times 320) \; + \; 64 \end{aligned}$$

Now rewrite, in reverse order, each of these divisions-with-remainder as an equation expressing the remainder as an integral linear combination of the divisor and dividend, thus:

$$\begin{aligned} 64 &= 1344 + ((-4) \times 320) \\ 320 &= 3008 + ((-2) \times 1344) \end{aligned}$$

Substitute the value of 320 in the second equation into the first, and we get:

$$\begin{aligned} 64 &= 1344 + (-4) \times (3008 + (-2) \times 1344) \\ &= 1344 + ((-4) \times 3008) + (8 \times 1344) \\ &= \{9 \times 1344\} + \{(-4) \times 3008\} \end{aligned}$$

In other words the GCD of 1344 and 3008 is 64, and

$$64 = \{9 \times 1344\} \ + \ \{(-4) \times 3008\}$$

Let us also express GCD$(10049, 1190)$ as an integral linear combination of 10049 and 1190. Again, we first list the steps of division with remainder:

$$\begin{aligned} 10049 &= (8 \times 1190) \ + \ 529 \\ 1190 &= (2 \times 529) \ + \ 132 \\ 529 &= (4 \times 132) \ + \ 1 \end{aligned}$$

As before, we rewrite each equation as an expression of the remainder in terms of the dividend and divisor, but in reverse order:

$$\begin{aligned} 1 &= 529 + ((-4) \times 132) \\ 132 &= 1190 + ((-2) \times 529) \\ 529 &= 10049 + ((-8) \times 1190) \end{aligned}$$

So we have, by repeated substitution:

$$\begin{aligned}
1 &= 529 + ((-4) \times 132) \\
&= 529 + (-4)(1190 + (-2) \times 529) \\
&= 529 + ((-4) \times 1190) + (8 \times 529) \\
&= (9 \times 529) + ((-4) \times 1190) \\
&= 9 \times (10049 + (-8) \times 1190) + ((-4) \times 1190) \\
&= (9 \times 10049) + ((-72) \times 1190) + ((-4) \times 1190) \\
&= \{9 \times 10049\} + \{(-76) \times 1190\}
\end{aligned}$$

Thus the GCD of 10049 and 1190 is 1, and

$$1 = \{9 \times 10049\} + \{(-76) \times 1190\}$$

Clearly, no one would consider this expression of 1 as an integral linear combination of 10049 and 1190 to be obvious.

The general case is quite clear at this point. Let whole numbers a and d be given, with $a > d$. If we iterate the process of performing division with remainder on a and d, and then on d and the remainder, etc., obtaining:

$$\begin{aligned}
a &= q\,d \ + \ r \\
d &= q_1\,r \ + \ r_1 \\
r &= q_2\,r_1 \ + \ r_2 \\
r_1 &= q_3\,r_2 \ + \ r_3 \\
r_2 &= q_4\,r_3 \ + \ 0
\end{aligned}$$

Note that the division with remainder can, in principle, continue for $d - 1$ steps before it terminates with remainder 0, but for simplicity of writing, we have allowed the remainder 0 to appear after 4 steps. Clearly there is no loss of generality in the reasoning. That said, we conclude that

$$\begin{aligned}
\text{GCD}(a, d) \ = \ \text{GCD}(d, r) \ = \ \text{GCD}(r, r_1) \ = \ \text{GCD}(r_1, r_2) \ = \\
\text{GCD}(r_2, r_3) \ = \ \text{GCD}(r_3, 0) \ = \ r_3.
\end{aligned}$$

Furthermore, the GCD, which is r_3, can be expressed as an integral linear combination of a and d as follows. First, rewrite the preceding sequence of equations in reverse order, each time expressing the remainder as a linear combination of the dividend and the divisor:

$$\begin{aligned}
r_3 &= r_1 + (-q_3)\, r_2 \\
r_2 &= r + (-q_2)\, r_1 \\
r_1 &= d + (-q_1)\, r \\
r &= a + (-q)\, d
\end{aligned}$$

Therefore, repeated substitution of r_i into the equation above it yields:

$$\begin{aligned}
r_3 &= r_1 \ + \ (-q_3)\, r_2 \\
&= r_1 \ + \ (-q_3)(r + (-q_2)\, r_1) \\
&= (-q_3)\, r \ + \ (1 + q_2\, q_3)\, r_1 \\
&= (-q_3)\, r \ + \ (1 + q_2\, q_3)\, (d + (-q_1)\, r) \\
&= (1 + q_2\, q_3)\, d \ + \ (-q_1 - q_3 - q_1\, q_2\, q_3)\, r \\
&= (1 + q_2\, q_3)\, d \ + \ (-q_1 - q_3 - q_1\, q_2\, q_3)\, (a + (-q)\, d) \\
&= (1 + q_2\, q_3 + q\, (q_1 + q_3 + q_1\, q_2\, q_3))\, d \ + \ (-q_1 - q_3 - q_1\, q_2\, q_3)\, a
\end{aligned}$$

We have therefore proved the following theorem:

Theorem 2 (Euclidean Algorithm) *Given $a, d \in \mathbf{N}$. Then* $\mathrm{GCD}(a, d)$ *can be obtained by a finite number of applications of division with remainder. Furthermore,* $\mathrm{GCD}(a, d)$ *is an integral linear combination of a and d.*

The reason we are interested in the Euclidean Algorithm in this context is that it leads directly to the

Proof of Key Lemma The following brilliant proof is (so far as we can determine) due to Euclid, which of course also accounts for the name. We are given whole numbers ℓ, m, and n, so that $\ell | mn$ and ℓ and m are relatively prime. We must prove $\ell | n$. Since ℓ and m are relatively prime, $\mathrm{GCD}(\ell, m) = 1$. By the Euclidean algorithm,

$1 = \alpha\ell + \beta m$ for some integers α and β. Multiply this equation through by n, and we get $n = \alpha\ell n + \beta mn$. Since ℓ divides mn by hypothesis, $\ell|(\beta mn)$; obviously, $\ell|(\ell n)$. Therefore ℓ divides $\alpha\ell n + \beta mn$, which is n. In other words, ℓ divides n. The proof is complete.

We are now in a position to give a proof of the main theorem of this section announced earlier: *Any fraction* $\frac{k}{\ell}$, *is equal to a unique fraction* $\frac{m}{n}$ *in reduced form. Moreover, there is an algorithm to produce this* $\frac{m}{n}$.

Proof of Main Theorem Let $\text{GCD}(k, \ell) = a$. Thus $k = am$ and $\ell = an$ for some whole numbers m and n. Note that m and n are relatively prime because if $\text{GCD}(m, n) = b > 1$, then $ab|(am)$ and $ab|(an)$, so that ab is a common divisor of k and ℓ which is bigger than a, contradicting the fact that a is the GCD of k and ℓ. Therefore $\frac{m}{n}$ is a reduced fraction. By equivalent fractions, $\frac{m}{n} = \frac{am}{an} = \frac{k}{\ell}$. Thus this $\frac{m}{n}$ is the desired fraction. Since $m = \frac{k}{a}$ and $n = \frac{\ell}{a}$, and since a is obtained from k and ℓ by the Euclidean algorithm, the theorem is proved with the exception of the uniqueness statement.

To prove uniqueness, suppose $\frac{k'}{\ell'} = \frac{k}{\ell}$ and $\frac{k'}{\ell'}$ is reduced. We must prove that $k' = m$ and $\ell' = n$. We have $\frac{k'}{\ell'} = \frac{m}{n}$, both being equal to $\frac{k}{\ell}$. By the cross-multiplication algorithm, $k'n = \ell'm$. Since $n|(k'n)$, we see that $n|(\ell'm)$. Since $\frac{m}{n}$ is reduced, m and n are relatively prime. Therefore the Key Lemma implies that $n|\ell'$, so that $n \leq \ell'$. We now look at $k'n = \ell'm$ from a different angle. Since $\ell'|(\ell'm)$, we have $\ell'|(k'n)$. Since $\frac{k'}{\ell'}$ is reduced, ℓ' and k' are relatively prime. By the Key Lemma, we must have $\ell'|n$ and thus $\ell' \leq n$. Together with $n \leq \ell'$, we get $n = \ell'$. Using $k'n = \ell'm$, we conclude that also $k' = m$, as desired.

Exercises 3.1

1. (i) If k, ℓ, m are integers, and if $k|\ell$ and $\ell|m$, then $k|m$. (ii) Every nonzero integer divides 0. (Caution: Use the precise definition of divisibility.)

2. Suppose A, B, C are whole numbers and $A = B + C$. If a whole number n divides any two of A, B, C, then n divides all three.

3. Find the GCD of each of the following pair of numbers by listing all the divisors of each number and compare: 35 and 84, 54 and 117, 104 and 195.

4. Find the GCD of each of the following pairs of numbers, and express it as an integral linear combination of the numbers in question: 322 and 159, 357 and 272, 671 and 2196.

5. Let the GCD of two positive integers a and d be k, and let $k = ma - nd$ for some whole numbers m and n. Then m and n are relatively prime.

6. In each of the following, find the reduced form of the fraction. (a) $\frac{160}{256}$. (b) $\frac{273}{156}$. (c) $\frac{144}{336}$. (d) $\frac{1147}{899}$.

7. The effectiveness of the Euclidean algorithm depends on how fast the remainders in the sequence of iterated divisions-with-remainder get to 0. Here is an indication: Suppose we have three iterated divisions-with-remainder as follows:

$$\begin{aligned} d &= q_1 r + r_1 \\ r &= q_2 r_1 + r_2 \\ r_1 &= q_3 r_2 + r_3 \end{aligned}$$

Then prove that $r_3 < \frac{1}{2} r_1$.

8. Prove that a whole number is divisible by 4 exactly when the number formed by its last two digits (i.e., its tens digit and ones digit) is divisible by 4. (Thus 93748 is divisible by 4 because 48 is divisible by 4.)

9. Prove that a whole number is divisible by 5 if and only if its last digit is 0 or 5.

10. The number 3 is a divisor of a whole number n if and only if 3 is a divisor of the number obtained by adding up all the digits of n. (*Hint*: 3 dividing a power of 10

always has remainder 1.)

11. Repeat problem 10 with the number 3 replaced by the number 9.

12. (a) For any whole number n, $\mathrm{GCD}(n, n+1) = 1$. (b) What is $\mathrm{GCD}(n, n+2)$ for a whole number n? (c) Let n be a whole number. What could $\mathrm{GCD}(n, n+k)$ be for a whole number k?

2 The fundamental theorem of arithmetic

The purpose of this section is to prove the following basic theorem and to use it to draw two consequences about numbers: the first about fractions which are equal to finite decimals, and the second about the existence of numbers (i.e., points on the number line) which lie outside **Q**.

Theorem 1 (Fundamental theorem of arithmetic) *Every whole number $n \geq 2$ is the product of a finite number of primes: $n = p_1 p_2 \cdots p_k$. Moreover, this collection of primes $p_1, \ldots, p_k$, counting the repetitions, is unique.*

This theorem will usually be referred to as **FTA**. The *uniqueness* statement, which is important for many reasons, can be made more explicit, as follows: Suppose $n = p_1 p_2 \cdots p_k = q_1 q_2 \cdots q_\ell$, where each of the p_i's and q_j's is a prime. Then $k = \ell$ and, *after renumbering the subscripts of the q's if necessary*, we have $p_i = q_i$ for all $i = 1, 2, \ldots, k$.

The expression of n as a product of primes, $n = p_1 p_2 \cdots p_k$, is called its **prime decomposition**. Let it be noted explicitly that in the above expression, some or all of the p_i's could be the same, e.g., $24 = 2 \times 2 \times 2 \times 3$. FTA says that, except for the order of the primes, the prime decomposition of each n is unique.

It should not be assumed that getting the explicit prime decomposition of a whole number is easy. Try 9167, for instance. Even with the help of the Lemma of the last section, we still have to check all the primes ≤ 96 to see if any of them divides 9167. It turns out that 9167 has the prime decomposition: $9167 = 89 \times 103$. The whole field of cryptography, which makes possible the secure transmission of confidential information on the internet, depends on the fact that if a number is very large, say 300 digits, then all the computers in the world put together cannot get its explicit prime decomposition in 100 years.

On the other hand, it is easy to establish that, *on a theoretical level*, every whole number has a prime decomposition.[38] Given $n \in \mathbf{N}$, if it is a prime, we are done. If not, n has a proper divisor. Among all its proper divisors, take the smallest, to be called p. Arguing as in the proof of the Lemma above, this p is a prime. Therefore let us write $n = pn_1$ for some whole number n_1. Apply the same argument to the whole number n_1, and we get $n_1 = qn_2$, where q, n_2 are whole numbers and q is a prime. Then we have $n = pqn_2$. Repeat the same argument on n_2, and after a finite number of steps, we get *a* prime decomposition of n.

It is the uniqueness that is more interesting and more difficult. The proof of uniqueness is mathematically sophisticated, and it is due to Euclid. Let us first convince ourselves that there is something to prove. Consider the following two expressions of 4410 as a product:

$$4410 = 2 \times 9 \times 245 = 42 \times 105$$

These two products, $2 \times 9 \times 245$ and 42×105, have different numbers of factors and the factors are all distinct. The *non*uniqueness of the expression of 4410 as a product is striking. Of course with the exception of 2, none of the factors is a prime. Once we require that each factor in the product be a prime, then we get only one possibility (other than those obtained by permuting the factors):

$$4410 = 2 \times 3 \times 3 \times 5 \times 7 \times 7$$

The question is: *why* must uniqueness emerge as soon as we require each factor to be a prime? The answer resides, in large part, in the Key Lemma of the last section.

Proof of uniqueness of prime decomposition Let n be a whole number so that $n = p_1 p_2 \cdots p_k = q_1 q_2 \cdots q_\ell$, where the p's and the q's are primes. We want to

[38] The difference between the explicit determination of a number and the theoretical statement that this number exists can be seen from an example. It is easy to write down a definite integral whose exact value is impossible to determine, e.g., $\int_0^7 \sin(x^{3.6})dx$, but the fact that this integral is equal to *some* number is easy to prove.

prove that $k = \ell$, and that after a renumbering of the subscripts of the q's if necessary, we have $p_i = q_i$ for all $i = 1, 2, \ldots, k$.

We first prove that p_1 is equal to a q_j, for some j. Since $p_1 | p_1 \cdots p_k$, we have $p_1 | q_1 \cdots q_\ell$. Write $Q_1 = q_2 \cdots q_\ell$, then $p_1 | q_1 Q_1$. If $p_1 = q_1$, we are done. If not, then p_1 and q_1 are distinct primes and are therefore relatively prime. By the Key Lemma of the last section, $p_1 | Q_1$. Writing Q_2 for $q_3 \cdots q_l$, we have $p_1 | q_2 Q_2$. Again, if $p_1 = q_2$, we are finished. If not, then p_1 and q_2, being distinct primes, are relatively prime. The Key Lemma implies $p_1 | Q_2$, etc. Either p_1 is equal to one of $q_3, \ldots q_{\ell-1}$, or after $\ell - 1$ steps, we have $p_1 | q_{\ell-1} q_\ell$. If $p_1 = q_{\ell-1}$, we are done. Otherwise, p_1 and $q_{\ell-1}$ are relatively prime and therefore $p_1 | q_\ell$. Since both p_1 and q_ℓ are primes, this is possible only if $p_1 = q_\ell$. Thus p_1 is equal to a q_j, for some j.

By relabeling the subscripts of the q's if necessary, we may assume that $p_1 = q_1$. The hypothesis that $n = p_1 p_2 \cdots p_k = q_1 q_2 \cdots q_\ell$ now reads: $p_1(p_2 \cdots p_k) = p_1(q_2 \cdots q_\ell)$. Multiply both sides by $(p_1)^{-1}$, and we get:

$$p_2 \cdots p_k = q_2 \cdots q_\ell$$

We may repeat the same argument, and conclude that p_2 is equal to one of $q_2, \ldots q_\ell$. By re-arranging the subscripts of the q's if necessary, we may assume that $p_2 = q_2$, so that

$$p_3 \cdots p_k = q_3 \cdots q_\ell,$$

etc. If $k \neq \ell$, let us say $k < \ell$, then after $k - 1$ such steps, we get

$$1 = q_{k+1} \cdots q_\ell$$

This is impossible because each of the primes $q_{k+1}, \ldots q_\ell$ is greater than 1. Thus $k = \ell$ after all, and $p_i = q_i$ for all $i = 1, \ldots k$. The proof of FTA is complete.

FTA has an interesting application to fractions. The following characterizes all the fractions which are equal to finite decimals.

Theorem 2 *If the denominator of a fraction is of the form $2^a 5^b$, where a and b are whole numbers, then it is equal to a finite decimal. Conversely, if a reduced fraction $\frac{m}{n}$ is equal to a finite decimal, then the prime decomposition of the denominator* *contains no primes other than 2 and 5.*

Note that the second part of the theorem is clearly false if $\frac{m}{n}$ is not reduced. For example, $\frac{3}{6} = 0.5$, but the prime decomposition of 6 contains a 3. It is also good to recall that a finite decimal is just a fraction whose denominator is a power of 10. It will be apparent from the proof how important it is to have such a clear-cut definition of a decimal.

We first prove that if the prime decomposition of the denominator n contains no primes other than 2 and 5, then $\frac{m}{n}$ is equal to a finite decimal. The idea of the proof is so simple that an example would give it away: since $160 = 2^5 \cdot 5$, the fraction $\frac{27}{160}$ is equal to the decimal 0.16875 because, by equivalent fractions,

$$\frac{27}{160} = \frac{27}{2^5 \cdot 5} = \frac{27 \cdot 5^4}{2^5 \cdot 5 \cdot 5^4} = \frac{16875}{10^5},$$

which by definition is 0.16875. In general, if $n = 2^a 5^b$, where a, b are whole numbers, we may assume without loss of generality that $a < b$. Then

$$\frac{m}{n} = \frac{m}{2^a 5^b} = \frac{2^{b-a} m}{2^{b-a} 2^a 5^b} = \frac{2^{b-a} m}{10^b}$$

and the last is a finite decimal.

Conversely, suppose $\frac{m}{n}$ is a reduced fraction which is equal to a finite decimal:

$$\frac{m}{n} = \frac{k}{10^b}$$

where k, b are whole numbers. We have to show that no prime other than 2 and 5 divides n. By the cross-multiplication algorithm, $nk = m10^b$. Since $\frac{m}{n}$ is reduced, n is relatively prime to m. Since n divides nk, it divides $m10^b$ as well. The Key Lemma shows that n must divide 10^b, which is $2^b 5^b$. Therefore, $2^b 5^b = n\ell$ for some whole number ℓ. By the uniqueness of the prime decomposition, the primes on the right consists of only 2's and 5's. Thus $n = 2^a 5^c$, where a and c are whole numbers $\leq b$. The theorem is proved.

At this point, we can take a look at the question of whether the rational numbers are sufficient for doing arithmetic. The following theorem implies that they are not, because many "square roots" cannot be rational numbers. For the statement of the theorem, a **perfect square** is a whole number which is equal to the square of another whole number. Thus the first few perfect squares are 1, 4, 9, 16, 25, 36, ...

Theorem 3 *If a whole number n is not a perfect square, then there is no rational number r so that $r^2 = n$.*

Proof Let the prime decomposition of n be expressed as a product of powers of *distinct* primes. (For example, $72 = 2^3\,3^2$, $3375 = 3^3\,5^3$, etc.) Consider the case where n is the product of powers of three distinct primes: $n = p_1^a\,p_2^b\,p_3^c$, where a, b, c are nonzero whole numbers, and p_1, p_2, p_3 are primes not equal to each other. The reasoning for this special case will be perfectly general, and by limiting ourselves to three primes, we save ourselves from some horrendous notation. If a, b, and c are all even, let $a = 2\alpha$, $b = 2\beta$ and $c = 2\gamma$ for some whole numbers α, β and γ. Then $n = (p_1^\alpha p_2^\beta p_3^\gamma)^2$, contradicting the hypothesis that n is not a perfect square. Therefore at least one of a, b, and c is odd, let us say, $a = 2k+1$ for some whole number k. Thus $n = p_1^{2k+1}\,p_2^b\,p_3^c$.

Suppose there is some rational number r so that $r^2 = n$. Let $r = \frac{A}{B}$, where A and B are whole numbers. Then

$$\frac{A^2}{B^2} = n = p_1^{2k+1}\,p_2^b\,p_3^c,$$

which implies that

$$A^2 = B^2\,p_1^{2k+1}\,p_2^b\,p_3^c.$$

By FTA, there are exactly the same number of p_1's on the left as on the right. We claim that the number of p_1's on the right is odd. Indeed, if the prime decomposition of B contains a p_1, then B^2 contains an even number of p_1's. There are of course $2k+1$ p_1's in p_1^{2k+1}, and there are no p_1's in either p_2^b or p_3^c because the three p's are distinct. Therefore there are an odd number of p_1's on the right, as claimed. But on the left, it is A^2. If there are j p_1's in the prime decomposition of A, where j is any whole number, then there are $2j$ p_1's in A^2. In any case, the number of p_1's on the left has to be even, which is a contradiction. Thus there is no such $r \in \mathbf{Q}$.

A number, i.e., a point on the number line, is formally called a **real number**. A real number is said to be **irrational** if it does not lie in $\mathbf{Q}$. At this point, we do not *know* if there is any irrational number or not because there is the possibility that every real number is rational, or equivalently, the whole number line is just $\mathbf{Q}$. While the preceding theorem says that the square root of many whole numbers is not a rational number, it does not by itself say that these square roots can be found among the real numbers. The fact that given any whole number n, *there is a real number* t so that t^2 makes sense and $t^2 = n$ will require a different kind of discussion, one that involves limits. This will be done in Chapter 16.

Exercises 3.2

1. Without using the Fundamental Theorem of Arithmetic, give a direct, self-contained proof of why the prime decomposition of 455 ($= 5 \times 7 \times 13$) is unique.

2. Given two positive integers a and b. If their GCD is k, then the two positive integers $\frac{a}{k}$ and $\frac{b}{k}$ are relatively prime.

3. Let a, b, c be positive integers. If a is relatively prime to b and both a and b divide c, then ab also divides c.

4. Define the **least common multiple (LCM)** of two whole numbers a and b to be the smallest whole number m so that m is a multiple of both a and b. (a) If $a = p^2q^7r^3$ and $b = p^6qs^4$, where p, q, r, s denote distinct primes, what are the GCD and LCM of a and b in terms of p, q, r, and s? (b) If k is the GCD of a and b, and m is their LCM, prove that $mk = ab$.

5. There are *consecutive odd numbers* which are primes, e.g., 5 and 7, 51 and 53, 101 and 103, etc. An example of three consecutive odd numbers which are primes is 3, 5, and 7. Are there other examples of three consecutive odd numbers which are primes?

6. A whole number which is the n-th power of another whole number is called a **perfect n-th power**. Prove that if a whole number k is not a perfect n-th power, there is no rational number whose n-th power is equal to k.

Chapter 4: Experimental Geometry

In this chapter, we begin an *informal* study of the geometry of the plane. It is a warm-up for the geometry discussion of the next three chapters. It outlines a series of suggested activities designed to foster the acquisition of geometric intuition. In the process, it also acquaints you with the working vocabulary and notations in geometry. Geometric intuition is important because, without it, not much geometry can be learned. We will not be overly concerned with total precision or total accuracy in this chapter; all the concepts introduced will be formally re-defined in subsequent chapters. The reason the issue of precision is relevant is that the requisite precision of formal definitions sometimes has the side effect of robbing a simple concept (such as the "direction of a translation") of its intuitive content. The purpose of this chapter is therefore to make sure you get acquainted with the underlying intuition before tackling the formal definitions.

1 Freehand drawing

Much geometric intuition can be developed by learning how to sketch common plane figures without the use of any tools (e.g., compass, ruler, protractor). This is the geometric analog of making estimates in numbers and operations. Eventually you will have to make precise drawings using, for example, a straightedge and compass, but before you get started, you should be able to see in your mind ahead of time roughly what kind of picture you are going to get. If you cannot do that, then your spatial intuition is under-developed and it will be difficult to engage in any serious geometric thinking. The ability to see a rough geometric picture without the use of any tools also has the benecial effect of averting gross mistakes, in much the same way that having an estimate of your answer ahead of time can prevent many mistakes due to pushing the wrong buttons on a calculator.

Here are some suggested drawings to be done in class:

1. Draw a line through a given point parallel to a given straight line.
2. Locate (approximately) the midpoint of each side of a given *obtuse* triangle. Use a ruler to check how good your guesses are.

3. Locate the midpoints of the three sides of a given triangle. Now join each vertex to the midpoint of the opposite side to obtain three segments. Do this for many triangles. *Question:* What do you notice about these segments?

4. Again draw a triangle and locate the midpoints of the three sides. This time, join these midpoints to each other get four smaller triangles within the original one. Do this for many triangles. *Question:* What do you notice about these smaller triangles?

5. Drop a perpendicular from a point outside a given line to the line. (The point where the perpendicular meets the line is called the **foot** of the perpendicular.)

6. Draw a line from a point on a given line L so that it is perpendicular to L.

7. Draw an acute angle, and drop a perpendicular from a point on one side of the angle to the (line containing the) other side. Do this for many angles. *Question:* Where does the perpendicular meet the other side? (An **angle** will be taken to be a *region* bounded by two **rays** (semi-infinite line segments), and each of these rays is called a **side** of the angle.)

8. Draw an obtuse angle, and drop a perpendicular from a point on one side of the angle to the (line containing the) other side. Do this for many angles. *Question:* Where does the perpendicular meet the other side?

9. Draw a triangle, and from each vertex, drop a line perpendicular to the (lines containing the) opposite side. Do this for many triangles, some obtuse and some acute. *Question:* What do you notice about these three perpendiculars?

10. Take a line segment, and draw the perpendicular through its midpoint. (The latter straight line is called the **perpendicular bisector** of the line segment.) *Question:* How is a point on the perpendicular bisector related to the two endpoints of the line segment?

11. Draw a triangle and draw the perpendicular bisector of each side. Do this for acute triangles as well as obtuse triangles. *Question:* What do you notice about these three perpendicular bisectors?

12. Draw a circle with a given center and a given radius.

13. Draw a circle passing through the three vertices of a given triangle. Try acute, obtuse, and right triangles. *Question:* What do you notice about the center of the circle?

14. Draw the angle bisector of a given angle. (The **angle bisector** is the ray that separates the angle into two equal angles.) From a point P on the angle bisector drop perpendiculars to both sides of the angle. *Question:* What do you notice about the the segments from P to the feet of the perpendiculars? Try drawing a circle with center at P and with one of these segments as radius. *Question:* What do you notice about this circle?

15. Draw the three angle bisectors of a given triangle. *Questions:* What do you notice about them? Can one draw a circle inside the triangle **tangent** to all three sides of the triangle (recall that *tangent to a side* means touching the side at exactly one point)?

Here are selected comments on the rationale behind the above suggested activities.

2. *The purpose is to develop an accurate perception of length, and the use of an obtuse triangle is to bring awareness of optical illusion. Even for people who get it right for acute triangles, they tend to trip themselves on the sides of the obtuse angle.*

3. *Can you see that these three lines* (*called the* **medians** *of the triangle*) *seem to meet at a point? This point is the* **centroid** *of the triangle.*

4. *Can you see that these four triangles are congruent to each other?*

7. and 8. *The two activities should acquaint you with the difference between acute and obtuse angles. In activity 8, the perpendicular does not meet the side as is; you have to extend the side to a line before the latter meets the perpendicular.*

9. *Again, you should get the idea that these perpendiculars probably always meet at a point, but in the case of an obtuse triangle, this point lies outside the triangle. This point is called the* **orthocenter** *of the given triangle. Let the triangle be ABC, and let its orthocenter be H. Now notice that if $\angle A$ is obtuse, then A becomes the orthocenter of the acute triangle HBC.*

10. *Observe the* **symmetry** *of the line segment with respect to its perpendicular bisector: every point on the latter is equi-distant from the endpoints.*

11. *The perpendicular bisectors meet at a point. Moreover, for obtuse triangles, this point is outside the triangle.*

12 *This will take practice, and some may never acquire the knack of drawing a good looking circle even with extended practice. One way to get around this is not to try to draw the circle in one fell swoop, but begin by putting a few points around the center which are judged to be at the same distance from the center as the given radius.*

13. *The center of this circle (called the* **circumcenter** *of the triangle) should be the point described in activity 11. One should also note the following: if the triangle is acute, its circumcenter is inside the triangle. If the triangle is obtuse, then its circumcenter is outside the triangle. But for a right triangle, the circumcenter should be the midpoint of the hypotenuse.*

14. *The angle should appear to be symmetric with respect to the bisector, in the sense that the two segments from P are equal in length. The circle in question should be tangent to the sides.*

15. *The angle bisectors meet at a point P, and the circle with center at P and with radius equal to any of the segments described in item 14 should be tangent to all three sides. This is called the* **incircle** *of the triangle. However, one should be careful not to confuse the point where the incircle touches a side with the point where the angle bisector of the opposite angle intersects the side.*

Exercises 4.1

All the problems ask for freehand drawings.

1. Given a quadrilateral. Can you draw a circle that passes through all four vertices? Do you know when this is possible?

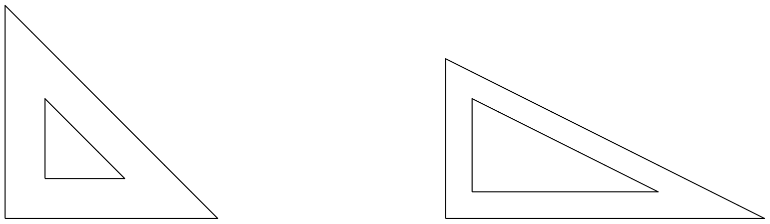
2. Let L_1 and L_2 be any two lines in the plane. Let points A, B, C be randomly chosen on L_1, and let points D, E, F be randomly chosen on L_2. Let AF, CD intersect at X, let AE, BD intersect at Y, and let BF, CE intersect at Z. What do you notice about X, Y, and Z?

3. Let a circle be given and let A be a point outside of this circle. Draw two lines through A; let the first intersect the circle at B and C and the second intersect it at D and E. Let CD and BE intersect at P, and BD and CE intersect at Q. Finally, let the line PQ intersect the circle at X and Y. What do you notice about the lines AX and AY?

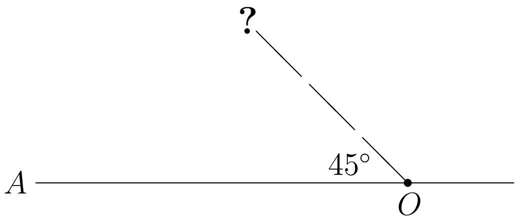
2 Constructions using tools

The use of plastic triangles

We next turn to some basic geometric constructions using tools. By tradition, i.e., by the tradition established by the Greek mathematicians around the fourth century BC, the tools of choice are ruler and compass (the "ruler" should be properly called a **straightedge** because we use it only to draw straight lines and never for measuring length, but this abuse of terminology is probably beyond recall). However, it must be said that there are a small number of constructions that are easier to do if we avail ourselves of **plastic triangles**, and we do these first. There are two kinds of plastic triangles on the market, the 90-45-45 one and the 90-60-30 one:

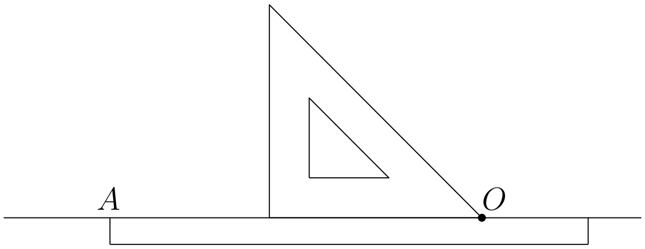


The simplest use of these triangles is ***to draw angles of 90, 60, 45, and 30 degrees.*** For example, to draw an angle of 45 degrees with a vertex O and one side OA given, we proceed as follows.



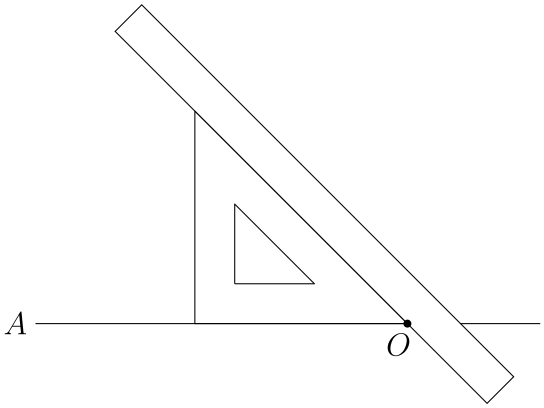
(1) Place a ruler along the line joining A and O, denoted by $\boldsymbol{L_{AO}}$, as shown.

(2) Holding the ruler firmly in place with one hand, place one **leg** of the 90-45-45 plastic triangle flush against the ruler and slide it until the 45-degree vertex of the triangle is at O, as shown.

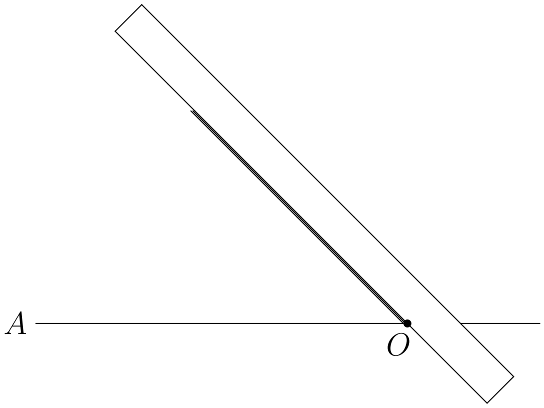


Now hold the triangle firmly in place and gently remove the ruler. If you draw the other side of the sought-after 45-degree angle at this point by running a pencil along the **hypotenuse** (i.e., the longest side) of the triangle, the line will not quite get to the point O neatly. Perform two more steps, as follows.

(3) *Press down with one hand on the plastic triangle to keep it in this position*, then use the other hand to place the ruler flush against the hypotenuse, as shown.



(4) *Press down on the ruler to keep it in this position* before gently removing the 90-45-45 triangle. Now draw the line from O that gives the other side of the desired 45-degree angle.

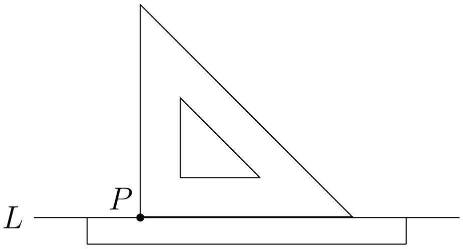


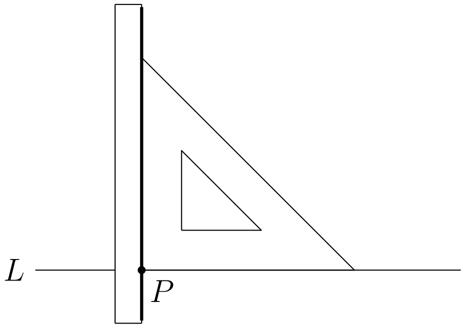
The next use of plastic triangles is ***to drop a perpendicular from a given point to a given line.*** For this purpose, either of the two plastic triangles can be used. For ease of drawing (on the computer), we will continue to use the 90-45-45 one. There are two possible scenarios: the point is on the given line, and the point is not on the given line.

First, suppose the point P lies on the given line L.

(1) Place a ruler along the line L, as shown in the left picture below.

(2) Put the triangle flush against the ruler and slide it along the ruler until the 90-degree vertex is at the point P, as shown in the left picture.





Once again, if you hold the plastic triangle firmly in place and run a pencil along the vertical side, you will get part of a line that doesn't quite reach P neatly. Do instead the following:

(3) With one hand holding the triangle firmly in place, place the ruler flush against the vertical side of the triangle as shown in the right picture above.

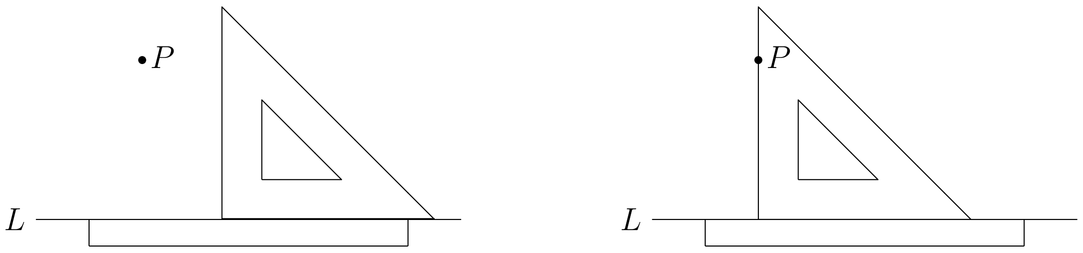
(4) Gently remove the triangle while holding the ruler in place. You can now draw a complete line passing through P and perpendicular to the line L.

Now consider the case of the point P not lying on L. Here are the steps.

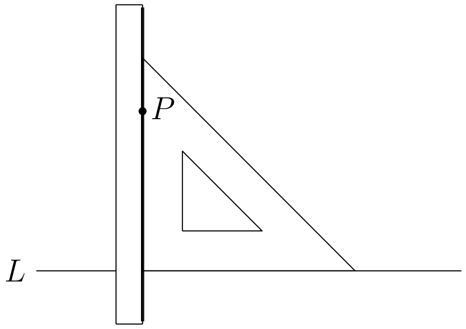
(1) Place a ruler along the line L as before.

(2) Pressing down on the ruler to hold it in place, put the triangle flush against the ruler, as in the picture on the left below.

(3) Still holding the ruler firmly in place, slide the triangle along the ruler until the vertical side of the triangle passes through P, as shown in the picture on the right.



By now, you are already experienced in this game and it should be unnecessary to repeat the steps about how to place a ruler flush against the vertical side of the triangle and draw the desired line along the ruler rather than along the vertical leg of the triangle.



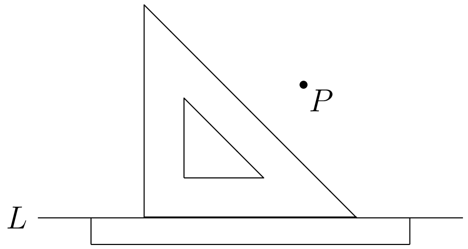
The final construction with plastic triangles is ***to draw a line through a given***

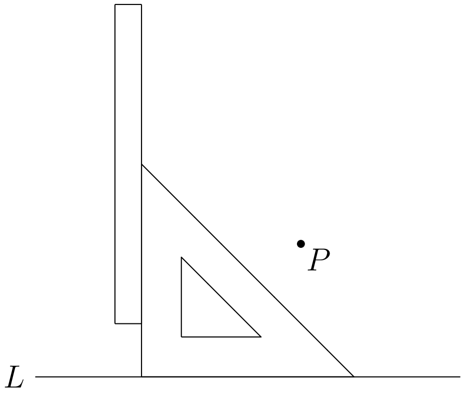
point P parallel to a given line L which does not contain P. Again, either triangle can be used, but we will continue to use the 90-45-45 triangle for illustration.

(1) Place a ruler along the line L.

(2) Holding the ruler in place with one hand, put the triangle flush against the ruler as shown in the left picture below.

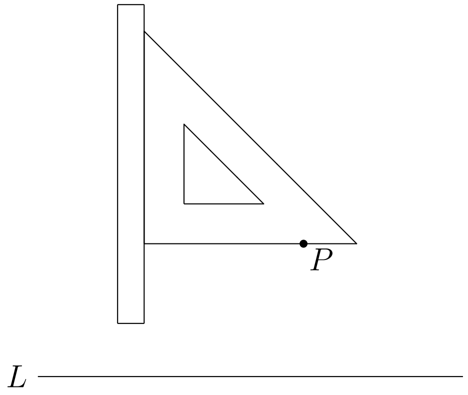
(3) Now hold the triangle firmly in place and put the ruler flush against the vertical side of the triangle, as shown in the right picture below.

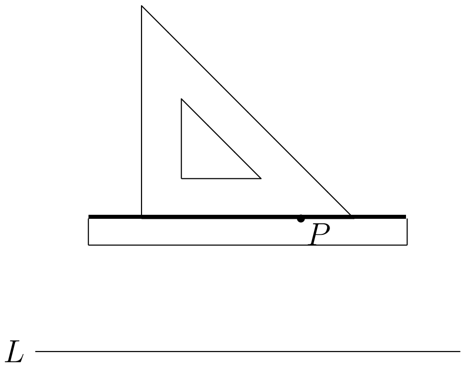




(4) Hold the *ruler* firmly in place and slide the *triangle* along the ruler until the horizontal side passes though the point P, as shown in the picture on the left below.

(5) Gently remove the ruler and draw the parallel line through P, or go through the standard procedure of placing the ruler flush against the horizontal side of the triangle before drawing the line, as shown in the picture on the right.





Remarks.

(a) One has to exercise good judgment about where to place the triangle at the beginning of the construction. Otherwise the critical step (4) cannot be carried out if the horizontal side of the triangle is too short to pass through P.

(b) We put one leg of the triangle flush against the ruler in step (2) above, but you can equally well do the same using the hypotenuse.

(c) This construction is most useful when we have to draw a line parallel to a given line L passing through each of several given points P_1, P_2, …, P_n. In this case, we would modify steps (4) and (5) as follows:

(4′) Hold the *ruler* firmly in place and slide the *triangle* along the ruler until the horizontal side passes though the point P_1. Draw the line through P_1.

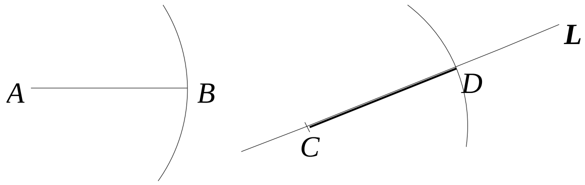
(5′) Still holding the *ruler* firmly in place, now slide the *triangle* along the ruler until the horizontal side passes though the point P_2. Draw the line through P_2. etc.

Using ruler and compass

We now turn to ruler-compass constructions. No proofs are necessary at this point. Most students derive a sense of power knowing how to use tools to get things done. If they can do this here, they will have a good entry into the world of geometry. We begin with the most basic constructions.

(a) Reproduce a line segment on another line with a specified endpoint.

Suppose a line segment AB is given; a line L is also given together with a point C on L. The problem is to construct a line segment on L which has C as an endpoint and which has the same length as AB.



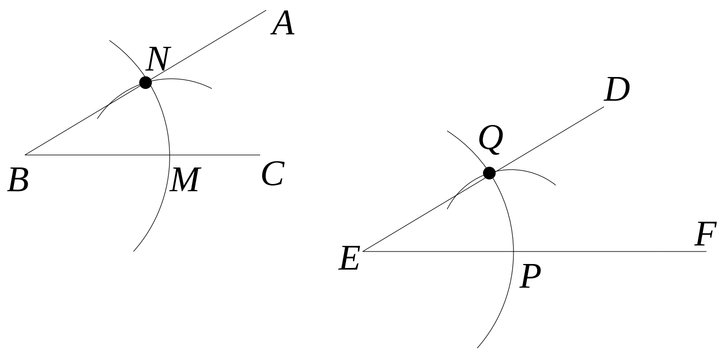
The construction:

1. Open the compass to the length of AB. With C as center and AB as radius, draw a circle which intersects L at two points. Let one of them be D.

2. CD is the desired segment.

(b) Reproduce an angle on a given ray.

Suppose we are given an angle $\angle ABC$ and a ray to be denoted by $\boldsymbol{R_{EF}}$. The problem is to construct a ray R_{ED} so that $\angle DEF$ equals (i.e., has same degree as) $\angle ABC$.



The construction:

1. Draw a circle (any radius) with B as center, then draw another circle with E as center but with the same radius.

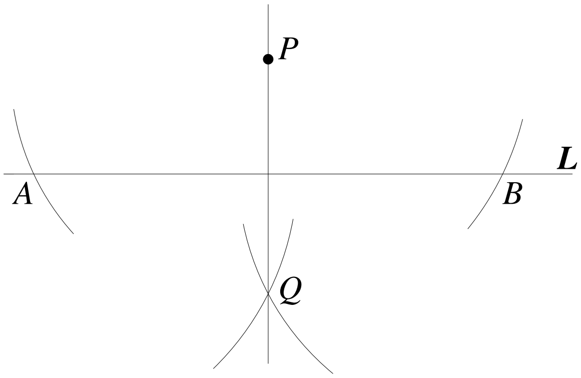
2. Let the first circle intersect the sides of $\angle ABC$ at N and M, as shown, and let the second circle (to be called *circle* E) intersect the ray R_{EF} at P.

3. With P as center and with the length of MN as radius, draw a circle and let one of the points of intersection of this circle with circle E be Q.

4. Let R_{ED} be the ray that contains E and Q. Then $\angle DEF$ is the sought-after angle.

(c) Construct a line from a point perpendicular to a line L.

Let the point be P. There are two cases to consider: P lies on L, and P does not lie on L. It will be seen that the following construction takes care of both cases at the same time.



The construction:

1. Draw a circle with P as center so that it intersects L at two points A and B.

2. Draw two circles with the same (sufficiently large) radius but with two different centers A and B so that they intersect; let one of the points of intersection be Q.

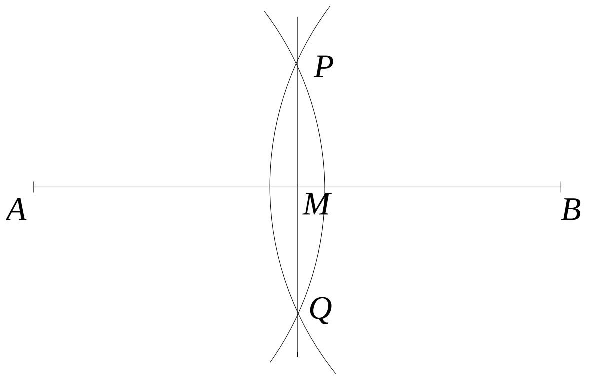
3. The line L_{PQ} is the line we seek.

(d) Construct the perpendicular bisector of a line segment.

Let the segment be AB. We have to construct a line which is perpendicular to AB and which passes through the midpoint of AB. Incidentally, *this construction also yields a method to locate the midpoint of a given segment.*

The construction:

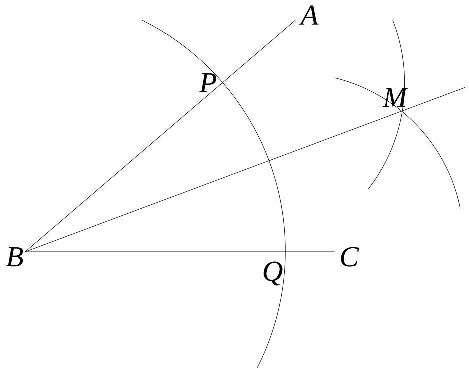
1. Draw two circles with the same (sufficiently large) radius but with two different centers A and B so that they intersect at P and Q.



2. The line containing P and Q is the line we seek. (By implication, the point of intersection M of PQ and AB is the midpoint of AB.)

(e) Construct the angle bisector of an angle.

Let the angle be $\angle ABC$. The problem is to construct a ray from B that bisects this angle. Note the similarity of this construction with construction (c).

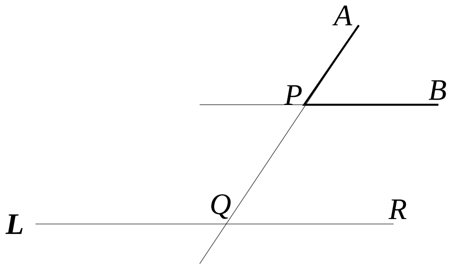


The construction:

1. Draw a circle with center at B (with any radius); let it intersect the rays R_{BA} and R_{BC} at P and Q, respectively.
2. Draw two circles with the same (sufficiently large) radius but with two different centers P and Q so that they intersect at a point M.
3. the ray R_{BM} is the angle bisector of $\angle ABC$.

(f) Construct a line through a point parallel to a line.

Let the line be L and let the point be P; P does not lie on L. We have to construct a line passing through P and parallel to L.



The construction:

1. Take a point Q on L and join P to Q.
2. Let A be a point on the ray R_{QP} so that P is between A and Q. Let R be some point on L. On the ray R_{PA}, construct an angle $\angle APB$ so that it is equal to $\angle PQR$, as shown in the picture. (See construction (b) above.)
3. The line passing through P and B is the line parallel to L.

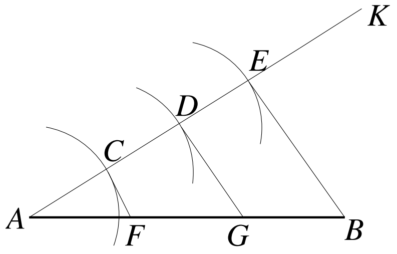
Remarks.

(*i*) This construction assumes that you are already completely fluent in reproducing an angle (construction (b)). If you experience difficulty in learning how to do this construction, please go back to construction (b) and learn that construction to the point of automaticity.

(*ii*) It is not easy to do this construction accurately, mainly because it is not easy to do construction (b) accurately. In practice, a better alternative is to use a plastic triangle to construct parallel lines.

(g) Divide a given line segment into any number of equal segments.

Let segment AB be given. We show how to trisect AB. The construction can obviously be generalized to equal division into any number of parts.



The construction:

1. Let R_{AK} be any ray issuing from A which is different from AB. Let AC be any segment on AK.

2. Reproduce AC successively on AK so that AC is equal to CD and is equal to DE (see (a)).

3. Join EB. From D and C, construct lines parallel to L_{EB} (see construction (f)), and let these lines intersect AB at G and F, respectively.

4. AF, FG, and GB have the same length.

Remark. Step (3) above can be done more accurately using a plastic triangle. See remark (*ii*) at the end of construction (f).

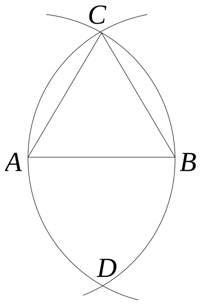
(h) Construct an equilateral triangle on a given side.

An **equilateral triangle** is a triangle whose sides have equal lengths. Let a segment AB be given. We have to construct an equilateral triangle with AB as one of its sides. (The picture is shown on the next page.)

The construction:

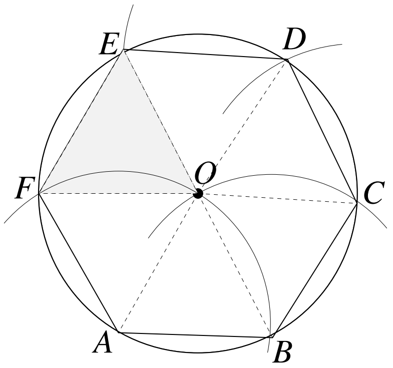
1. Draw two circles with AB as radius but with different centers at A and B.

2. Let C and D be the points of intersection of the two circles. Then $\triangle ABC$ or $\triangle ABD$ is the sought after triangle.



(i) Construct a regular hexagon in a circle.

A **regular hexagon in a circle** is a polygon whose six vertices lie on the circle, so that its six sides are equal and its six angles at the vertices are also equal. Given a circle with center O, we will refer to it as **circle O**. The problem is to locate six points on circle O so that they form the vertices of a regular hexagon.



The construction:

1. Take a point A on circle O whose radius will be denoted by r. With A as center and with r as radius, draw a circle which intersects circle O at B and F.

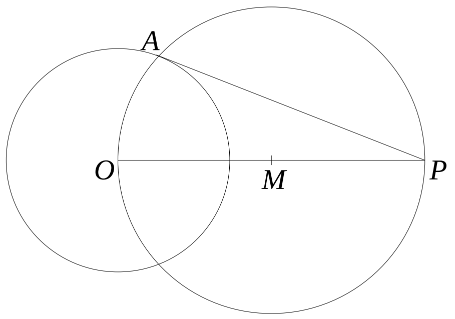
2. With B as center and r as radius, draw a circle which intersects circle O at an additional point C.

3. Repeat the drawing of circles with center C and then D, as shown, so that we obtain two more points D, and E.

4. Connect the successive points A, B, ..., F and A to get the desired hexagon.

(j) Draw tangents to a circle from a point outside the circle.

Let P be a point outside circle O. The problem is to construct a line passing through A and tangent to circle O.



The construction:

1. Join P to O to obtain segment OP.
2. Locate the midpoint M of OP (see construction (d)).
3. With M as center and MP as radius, draw a circle that intersects circle O at two points.
4. If A is a point of intersection in step 3, then the line L_{PA} is tangent to circle O.

Exercises 4.2

1. Accepting the fact that all the constructions with ruler and compass are mathematically correct, describe how you would go about constructing a square when one side is given. Discuss at which point of the construction you may have doubts that you are getting a square.

2. If you accept that the construction described in (f) is correct, explain on this basis why the earlier construction of a parallel line to a given line using plastic triangles is

correct.

3. (a) Perform construction 3 of §1 using ruler and compass. (b) Do the same with construction 9 of §1. (c) Do the same with construction 11 of §1.

4. Divide a given segment into 7 parts of equal lengths.

5. It is known that if A is a point on a circle centered at O and a line ℓ passes through A, then ℓ is tangent to the circle if and only if the line L_{OA} is perpendicular to ℓ. On this basis, (a) if PQ is a **diameter** of a circle (i.e., PQ passes through the center of the circle) and K is a point on the circle, what would you guess is the degree of $\angle PKQ$, and can you explain your guess? (Look at construction (j).) (b) If A is a point on a circle, use ruler and compass to draw a tangent to the circle at the point A.

3 The basic isometries, Part I

Translation along a vector

Reflection across a line

Rotation around a point

We now embark on some activities that are designed to familiarize you with a concept that is not only basic in the remainder of these notes, but basic in geometry itself: **isometry**[39] in the plane. This is a "rule" or a "motion" that specifies how to move each point in the plane to another point in the plane, in such a way that if two points are distance d apart, then they are still exactly distance d apart after being moved. For this reason, an isometry is sometimes called a **rigid motion**. In technical terms, an isometry (in the plane) is a motion that is **distance preserving**. With one exception, we have no example of isometries in the plane at this point, i.e., we know

[39] The prefix "iso" means *equal*, and the suffix "metry" means roughly *the process of measuring* (as in geo*metry*). This word deserves to be used in school mathematics.

of no rule that is distance preserving, and the exception is the rule that sends every point p to p itself. However, we had better come up with some interesting examples, and *fast*.

We introduce three basic rules, called **basic isometries**:

translation along a vector

reflection across a line

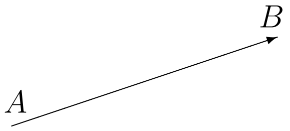
rotation around a point

We first explain what these rules are, and then we will convince ourselves through some hands-on activities that, indeed, they are distance preserving, i.e., they are isometries.

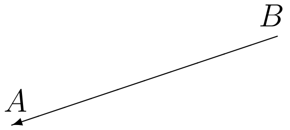
An easy way to come to grips with these three basic isometries is to get hold of a piece of paper and a sheet of overhead-projector transparency. We will imagine that the piece of paper is the plane and then we move the transparency in a way that simulates the effect of the basic isometry in question. A case-by-case discussion follows.

Translation along a vector

Translation is conceptually the simplest of the three basic isometries, although it will turn out to be difficult to make precise mathematical sense of the concept of the "direction" of a translation. Here we are more concerned with intuitive understanding and are willing to sacrifice precision for this purpose. First we generalize the concept of a **vector**, which was already used in §2 of Chapter 2. By a vector $\overline{AB}$, we mean a line segment AB together with the designation of the first point A as the **beginning point** and the second point B as the **endpoint**. Therefore $\overline{AB}$ and $\overline{BA}$ are different vectors although they have the same segment in common. We usually use an arrow at the endpoint to indicate, for example, that $\overline{AB}$ is *from A to B*. Thus if $\overline{AB}$ is the following vector,



then $\overline{BA}$ would be pictured as:



With a vector $\overline{AB}$ given, we define the **translation along $\overline{AB}$**, or more simply the **translation from A to B**, as the rule which moves points in the plane as follows: a point P in the plane is moved to the point Q so that

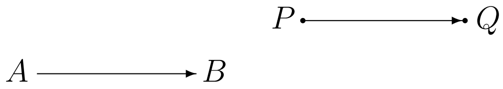
(i) if P does not lie on the line L_{AB} containing AB, then the (line containing the) segment PQ is parallel to the (line containing the) segment AB,

(ii) If P lies on L_{AB}, then Q is *in the same direction* relative to P as B is to A,

(iii) PQ has the same length as AB, and

(iv) the two vectors $\overline{PQ}$ and $\overline{AB}$ point *in the same direction*.

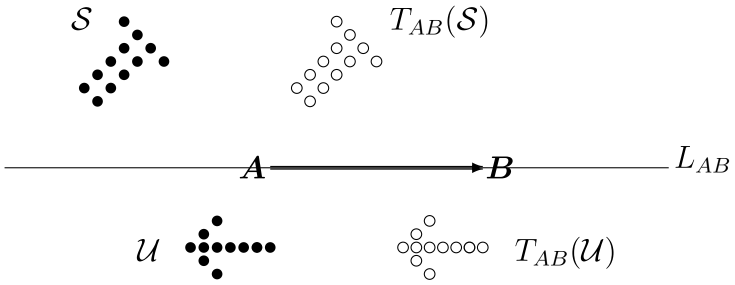
We leave open the precise meaning of "in the same direction" for now, but trust that on an intuitive level its meaning is unambiguous. For example, suppose AB is horizontal and points right, then $\overline{PQ}$ will likewise point right (and of course PQ is also horizontal because of (i)).



We usually denote a translation by T, or if necessary, by $\boldsymbol{T_{AB}}$ to denote the translation from A to B. A translation is easily visualized through the following activity.

Activity We are going to use a piece of paper as a model for the plane. On the paper, draw a vector AB, and also extend the segment AB to a line, denoted as usual by L_{AB}. Draw some figures on the paper. Then use a piece of overhead-projector transparency to copy everything on the paper, using (let us say) a red pen. In particular, both the vector $\overline{AB}$ and the line L_{AB} are on the transparency. Holding the paper in place, slide the transparency along the line L_{AB} until the *red* point A on the transparency is on top of the point B on the paper. The new positions of all the red figures on the transparency then display how the translation from A to B moves the figures on the paper.

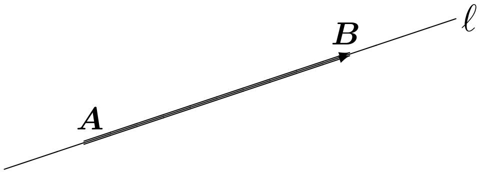
If AB is horizontal, the following is an example of what a translation does. We use the following notation: $T_{AB}(\mathcal{S})$ denotes the new position of a figure $\mathcal{S}$ after being moved by T_{AB}.



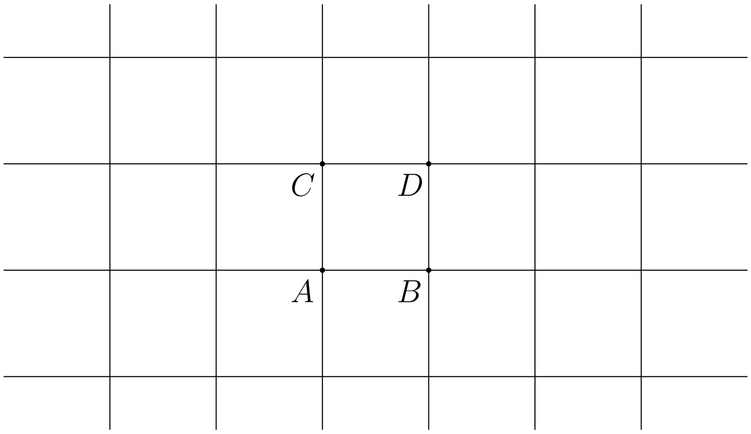
Can you explain why this activity is an accurate model of the concept of a translation?

We will formalize some concepts. Let T be the translation from some point A to some B. If T moves a point P to a point Q, then we denote Q also by $T(P)$ and call it the **image of P under T**, or more simply the **translated image of P**. We also say T **maps P to Q**. If $\mathcal{S}$ is a geometric figure in the plane, then the totality of all the *translated images* of the points in $\mathcal{S}$ is called the **image of $\mathcal{S}$ under T**, or more simply the **translated image of $\mathcal{S}$**, and denote it by $\boldsymbol{T(\mathcal{S})}$ (note the consistency with the notation in the preceding picture).

A geometric figure $\mathcal{S}$ is said to has **translational symmetry** if there is a translation T so that **T maps $\mathcal{S}$ onto itself**, i.e., $T(\mathcal{S}) = \mathcal{S}$, i.e., $\mathcal{S}$ is its own translated image. Note that for a figure $\mathcal{S}$ to possess translational symmetry, all it takes is to get one translation T so that $T(\mathcal{S}) = \mathcal{S}$. In practice, most figures with a translational symmetry possesses more than one translation to fulfill this requirement. The simplest figure with translational symmetry is a line ℓ: just pick *any* two points A and B on ℓ and you should be able to show that $T_{AB}(\ell) = \ell$:



Another obvious figure with translational symmetry is the grid on graph paper. With the points A, B, C, and D as indicated, if T denotes the translation along any of the three vectors $\overline{AB}$, $\overline{AC}$, and $\overline{AD}$ and if $\mathcal{G}$ denoted the grid, then $T(\mathcal{G}) = \mathcal{G}$.



Most patterns on rugs have translational symmetry, but the most famous geometric figures with translational symmetry are no doubt some of the prints of M. C. Escher. If you go to

http://www.mcescher.com/Shopmain/ShopEU/facsilimeprints/prints.html

you will get to see five of them: Two Birds, Bird/Fish, Lizard, Horseman, and Twelve Birds.

In view of the preceding Activity, the following assertions should be completely plausible for a translation T:

(T1) A translation is an isometry (i.e., distance-preserving); it also preserves the degree of an angle.

(T2) The image of a (straight) line under a translation is a (straight) line. Furthermore, the image of a segment is a segment, and the image of a ray is a ray.

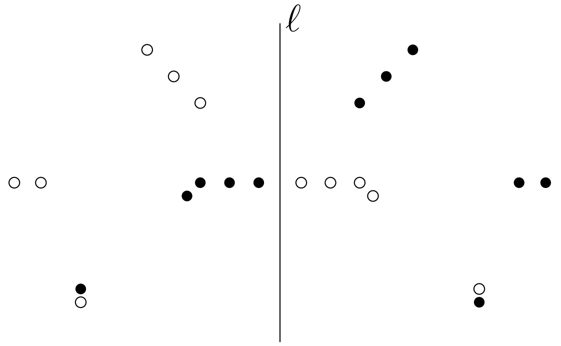
(T3) If T is a translation from A to B, then the distance between a point P and its translated image under T is always equal to the length of AB.

Some other simple properties of translation will be left to an exercise.

Reflection across a line

Given a line ℓ in the plane. We are going to describe a rule that moves points from one side of ℓ to the other side. We will use the letter R to denote this rule, to be called the **reflection across ℓ** or **reflection with respect to ℓ**: if p is a point in the plane that does not lie on ℓ, then R moves p to another point, traditionally denoted by $R(p)$, so that ℓ is the perpendicular bisector of the segment joining P to $R(p)$. Clearly, p and $R(p)$ lie on opposite sides of ℓ. On the other hand, if p is a point on ℓ itself, then by definition, $R(p) = p$, i.e., we leave p alone. We call $R(p)$ the **image of p under R**, or more simply as the **reflected image** of p. Thus, for *each and every point* in the plane, we have described how to move p to another point $R(p)$.

We shall describe the rule $p \mapsto R(p)$ in a slightly different language presently, but first, if ℓ is a vertical line, then we can easily picture p and $R(p)$ this way. Let us denote the reflected image of a point $\bullet$ by $\circ$, i.e., for the moment, we will use the representation that $R(\bullet) = \circ$. Then we have the following examples when ℓ is vertical:

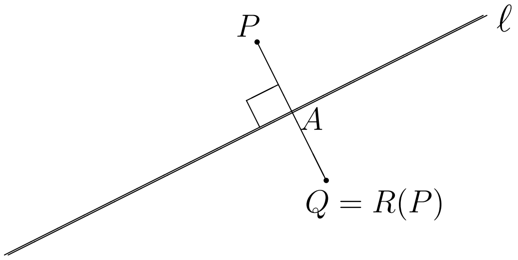


Note especially the fact that if the number line is inserted as a horizontal line in this picture, then for every p on the number line, the point $R(p)$ is exactly what we called p^* in §1 of Chapter 2. (This explains the terminology of "mirror reflection" of Chapter 2.) Note also that, just as $p^{**} = p$ in Chapter 2, we have in the case

$$R(R(p)) \;=\; p \quad \textit{for every point } p \textit{ in the plane}$$

where the meaning of $R(R(p))$ is as follows. Suppose q stands for $R(p)$, then q is a point in the plane and therefore $R(q)$ makes sense. So this $R(q)$ is exactly the meaning of $R(R(p))$. In the above picture, the meaning of this equality is that if we start with a point $\bullet$, the its reflected image is the corresponding $\circ$ on the opposite side of ℓ, but if we now look at the reflected image of this $\circ$, then it is back to the original $\bullet$.

We can define the reflection R in a different but equivalent way. Given ℓ and a point P not lying on ℓ, let the perpendicular line from P to ℓ meet ℓ at A (i.e., A is the foot of the perpendicular from P to ℓ). If Q is the point on this perpendicular, but on the other side of ℓ, so that the segment from P to A and the segment from A to Q have the same length, then $R(P) \;=\; Q$.



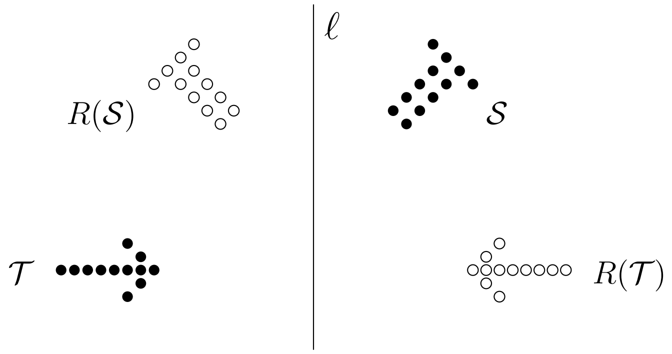
There is clearly no mathematical difference between this description of R and the original description, but this description has the advantage of making it easier to picture what R does: go straight from P to ℓ, and then go again the same distance and stop, and the stopping point is $R(P)$

Next, we give an intuitive description of R that gives a *global* picture of what R does. Go back to the picture of the vertical line above: one can imagine that if we fold the page across the vertical line, then each black dot $\bullet$ would fall on the corresponding white dot $\circ$ and vice versa. So a reflection is nothing but a precise way of describing how points are matched up when we fold the paper along the line ℓ if one can imagine the whole plane as a piece of paper. This then suggests an activity, one that is strongly recommended for the middle school classroom to acquaint students with the concept of a reflection.

Activity On a piece of paper, draw a line, to be called ℓ for the sake of discussion. Draw some figures on the paper. Then use a piece of overhead-projector transparency to *carefully* copy everything that is on the paper, *using a different color*, say red. In particular, the line ℓ is also on the transparency. Flip over the transparency and superimpose it on the paper, making sure that the red line ℓ on the transparency matches point-for-point the line ℓ on the paper. Now a comparison between the figures on the paper and the corresponding red figures on the transparency gives a clear idea of how the reflection across ℓ moves the points around: the red version of a figure on the paper is the reflection of the corresponding figure across ℓ.

Discuss with your neighbors why flipping the transparency over, as above, is an accurate realization of the reflection across ℓ.

The preceding activity highlights the need to consider, not just how each point is individually reflected, but how a whole geometric figure is reflected. Let R be the reflection across a line ℓ as usual. If $\mathcal{S}$ is a geometric figure in the plane, then the totality of all the *reflected images* of the points in $\mathcal{S}$ is called the **reflected image of $\mathcal{S}$**, or more generally the **image of $\mathcal{S}$ under R**, and is denoted $\boldsymbol{R(\mathcal{S})}$. Here are two examples in case the line ℓ is the vertical line. Note the effect of a reflection on "left" and "right", and on "up" and "down".



We now make several observations about reflections, which are all pictorially plausible, and are made all the more so by the preceding activity with transparencies. Let

R be a reflection across a line ℓ. Then:

(R1) A reflection is an isometry (i.e., distance-preserving) and, in addition, preserves degree of angles.

(R2) The image of a (straight) line under a reflection is a (straight) line. Furthermore, the image of a segment is a segment, and the image of a ray is a ray.

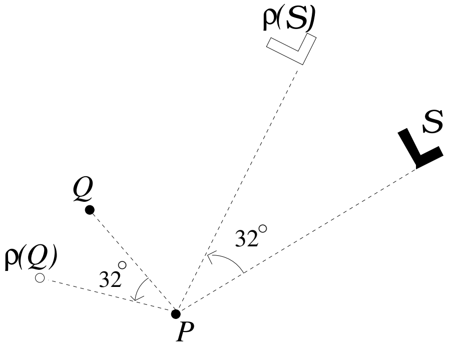
(R3) Reflecting twice across the same line leaves every point in the plane **fixed**, i.e., unchanged.

If the image of a geometric figure under R is the geometric figure itself, then we say ***R* maps the figure onto itself**. In school mathematics, it is more common to say that the figure has **bilateral symmetry** or **reflection symmetry** and ℓ is called the **axis of symmetry** or **line of symmetry**. For example, the letters "A", "H", "M" (among others) have bilateral symmetry with respect to the vertical line in the middle, whereas a circle has bilateral symmetry with respect to *every* line passing through the center. One would like to believe that a photograph of every human face has bilateral symmetry with respect to the vertical line in the middle, but that is just wishful thinking. We leave as an exercise to show that, if we believe in (R1) – (R3), then the angle bisector of an angle is a line of symmetry of the angle and the perpendicular bisector of a segment is a line of symmetry of the segment.

Rotation around a point

Given a point P, we will define what is meant by *a rotation of* (let us say) *32 degrees around P*. Denote this rotation by ρ (Greek letter *rho*; we cannot use "R" because this letter has been preempted by reflections). The point P will be called the **center of rotation** and the number 32 the **degree of rotation**. In this case, the definition will be best explained by an activity.

Activity On a piece paper, fix a point P, and then draw a geometric figure $\mathcal{S}$. On a piece of transparency, copy all this information exactly, in red color (say), and keep the transparency in exactly this position. In particular, the red point P on the transparency is on top of the point P on the paper. Now use a pointed object (e.g., the needle of a compass) to pin the transparency to the paper *at the point* P. Holding the paper fixed, rotate the transparency around P, counterclockwise[40] by 32 degrees and stop; the position of the red figure is exactly where ρ moves $\mathcal{S}$. If the figure consists of a single point Q, then the position of the red Q is where the rotation ρ moves Q. Notice that ρ does not move P, the center of rotation. See picture; we will explain the notations used in the picture below.



Of course there is nothing special about the number 32. So we may define a **rotation with center P of degree e**, where $-180 \leq e \leq 180$, to be the rule which leaves P itself unchanged but which moves every point Q which is not P in the following way:

> If $e \geq 0$, join Q to P and rotate the segment PQ, with P as pivot, like the hands of a clock e degrees *counterclockwise*. The new position of Q is where ρ moves Q, and is denoted by $\rho(Q)$. If $e < 0$, then do the same except that we rotate PQ *clockwise* e degrees.

The point $\rho(Q)$ is called the **image of Q under ρ**, or more simply, the **rotated image of Q**. If $\mathcal{S}$ is a geometric figure, then the collection of all the rotated images of the points in $\mathcal{S}$ is called the **image of $\mathcal{S}$ under ρ**, or the **rotated image of $\mathcal{S}$**, to be denoted by $\rho(\mathcal{S})$. Compare the preceding picture.

It follows from the Activity that if ρ is a rotation of e degrees around a point P, and ρ' is a rotation of $-e$ degrees *around the same point* P, then rotating a point Q

[40] If "32 degrees" is replaced by a negative degree, then we rotate *clockwise*.

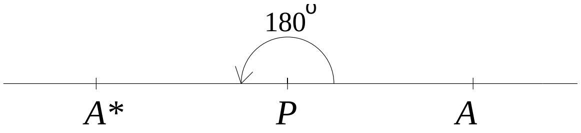
first by ρ and then by ρ' bring Q back to itself, i.e., leaves Q fixed. The succession of these two rotations is recorded as $\rho'(\rho(Q))$; let us make sure that this notation makes sense, and it does because if we write Q' for $\rho(Q)$, then Q' is just a point in the plane and therefore $\rho'(Q')$ makes sense. So this is exactly $\rho'(\rho(Q))$. Notice that, more generally, if ρ^* is a rotation of degree e^* *around the same point* P, then the rotation ρ followed by the rotation ρ^* is a rotation of degree $e + e^*$ around P. This is easily borne out by rotating the transparency, first e degrees, and then e^* degrees (keeping in mind that a rotation of positive degree is counterclockwise and a rotation of negative degree is clockwise.

For a later need, we want to address two issues concerning rotations. The first is: what does a rotation do to lines and angles? The preceding activity should convince you that, under a rotation, the image of a line is a line, the image of a segment of length d is a segment of length d, and the image of an angle is an angle of the same degree. Thus if $\angle ABC$ has degree e, then a rotation ρ, regardless of what its center is or what its degree of rotation may be, will rotate it to an angle $\rho(\angle ABC)$ of the same degree e. The second question is about rotations of 180 degrees. Let ρ^* be such a rotation, with center P, and let ℓ be a line not containing P. Take a point A on ℓ. There are two questions:

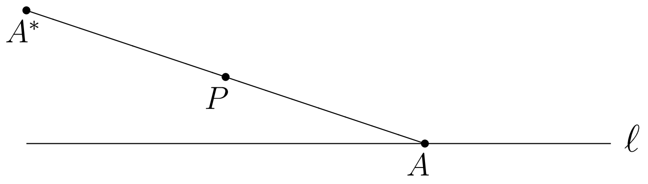
(*i*) Is there anything special about the three points A, P, and $\rho^*(A)$?

(*ii*) Can $\rho^*(A)$ be a point on ℓ, given that A is a point on ℓ?

The answer to (*i*) is ***yes***, and the reason is straightforward: write A^* for $\rho^*(A)$, then the fact that the angle $\angle APA^*$ is 180 degrees means that the the three points are **collinear**, i.e., lie on a line:



The answer to (*ii*) is more tantalizing. Again write A^* for $\rho^*(A)$, then the following picture shows clearly that A^* is "away" from ℓ, because it is as far on one side of P and A is on the *other* side. Therefore, A^* *cannot* lie on ℓ.



We will discover, gradually, that an indispensable part of learning geometry is to learn how to replace intuitive, gut feelings by mathematical reasoning. In this case, since we are completely convinced that A^* cannot be on ℓ, why not pretend that A^* is on ℓ and see what absurdity this assumption leads to? So let $A^* \in \ell$, then you'd notice this peculiarity about the line L_{AA^*}: the two lines ℓ and L_{AA^*} both join the point A to the point A^* which is different from A (*why are A and A^* different?*). But we know that there is only one line joining two distinct points, so we have to conclude that ℓ and L_{AA^*} coincide. *So what is wrong with that?* Here is where we must remember what we started with, i.e., what our *assumptions* are. We have assumed from the beginning that ℓ does not contain P. But L_{AA^*} contains P and if it coincides with ℓ, then ℓ must contain P. In the face of this contradiction, we see the error of our ways: we made the mistake of saying A^* is on ℓ. Therefore A^* is *not* on ℓ. Thus the answer to (*ii*) is ***no***.

Let us summarize our reasoning into one conclusion:

> *($\sharp$) If P is a point not on a line ℓ, and if ρ^* is the rotation of 180 degrees around P, then for every point A on ℓ, the rotated image $\rho^*(A)$ does not lie on ℓ.*

We will have many occasions to revisit this seemingly innocuous statement ($\sharp$).

As before, if a geometric figure $\mathcal{S}$ satisfies $\rho(\mathcal{S}) = \mathcal{S}$ for some rotation ρ, then we say $\mathcal{S}$ has **rotational symmetry**. We also say $\boldsymbol{\rho}$ **maps** $\boldsymbol{\mathcal{S}}$ **onto itself**. The prototypical figure with rotational symmetry is of course the circle: if O is the center of a circle, then the rotation of *every* degree with center at O maps the circle onto itself. We cannot *prove* it now, but you can believe that the regular hexagon of construction (i) in the preceding section has rotational symmetry: the rotation of 60 degrees around the center of the circle maps the hexagon onto itself. One of Escher's most famous prints, Circle Limit III, has a subtle 180 degree rotational symmetry. See http://www.d.umn.edu/~ddunham/dunamtlk.pdf

Just as with translations and reflections, several observations about rotations are entirely believable.

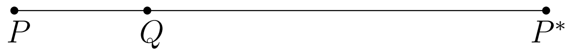
($\boldsymbol{\rho}$1) A rotation is an isometry and it preserves degrees of angles.

($\boldsymbol{\rho}$2) The rotated image of a line is a line, the rotated image of a segment is a segment, and the rotated image of a ray is a ray.

($\boldsymbol{\rho}$3) If ρ and ρ^* are two rotations — of degrees e and e^*, respectively, — with the same center P, then the point $Q' = \rho^*(\rho(Q))$ obtained by first rotating a point Q by ρ and then by ρ^* is the same point as the one obtained by rotating Q $e + e^*$ degrees around P.

In connection with **($\boldsymbol{\rho}$3)**, it is important to remember that each rotation depends on *two* pieces of data: the degree of the angle of rotation and the center of rotation. Whenever you talk about a rotation, *do not assume that the center of rotation is somehow understood* and pay attention only to the degree of the rotation. You must know what the center of rotation is! Here is an activity to remind you of this fact.

Activity In the picture below, let ρ, ρ^* be respectively the rotation of 90 degrees with center P, P^*, and Q is a point on the segment PP^* so that the length of QP^* is 3 times the length of PQ. Compare the following four points: $\rho^*(\rho(Q))$, $\rho(\rho^*(Q))$, the point A obtained by rotating Q 180 degrees around P, and the point A^* obtained by rotating Q 180 degrees around P^*.



Exercises 4.3

1. Explain why equilateral triangles, squares, and regular hexagons all possess rotational symmetry. Do they possess bilateral symmetry too? How many of the latter

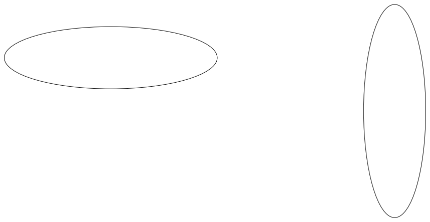
are there in each case?

2. Show that if ℓ is the angle bisector of an angle, then the reflection R across ℓ interchanges the two sides of the angle, i.e., if OA and OB are the sides of the angle, then $R(OA) = OB$ and $R(OB) = OA$.

3. Let T be the translation from point A to point B. Prove that if a line ℓ is neither L_{AB} nor parallel to L_{AB}, then the image of ℓ under T is parallel to ℓ.

4. Repeat the Activity above Exercises 4.3, but change ρ, ρ^* to be, respectively, the rotation of 30 degrees with center P, P^*. Use a protractor to make the drawings as accurate as possible.

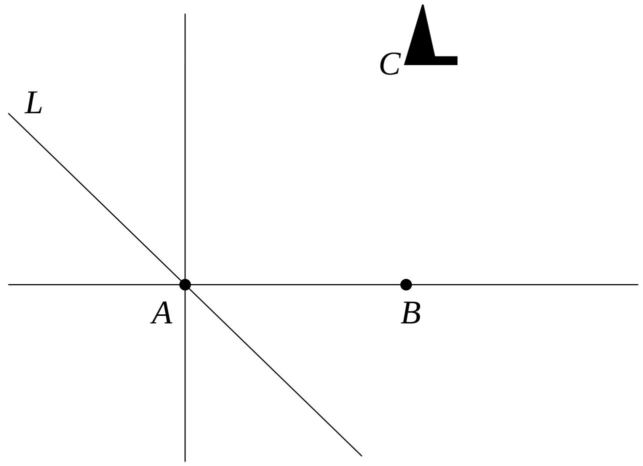
5. Describe a sequence of basic isometries that would move the left ellipse to the right one (there is obviously more than one way):



6. In the picture below, C denotes the lower left corner of the black figure, $|\angle CAB| = 45°$, $|AB| = |BC|$, and line L makes 45 degrees with line L_{AB}.

Let G be the clockwise rotation of $90°$ with center at the point A, let H be the reflection across the line L, and let J be the translation along $\overline{AB}$. Furthermore, let $\mathcal{S}$ denote the black figure.

Using a separate sketch for each of the following items, indicate the positions of (a) $J(\mathcal{S})$ and $G(J(\mathcal{S}))$, (b) $G(\mathcal{S})$ and $J(G(\mathcal{S}))$, (c) $H(\mathcal{S})$ and $G(H(\mathcal{S}))$, (d) $G(\mathcal{S})$ and $H(G(\mathcal{S}))$, (e) $J(\mathcal{S})$ and $H(J(\mathcal{S}))$, (f) $H(\mathcal{S})$ and $J(H(\mathcal{S}))$.



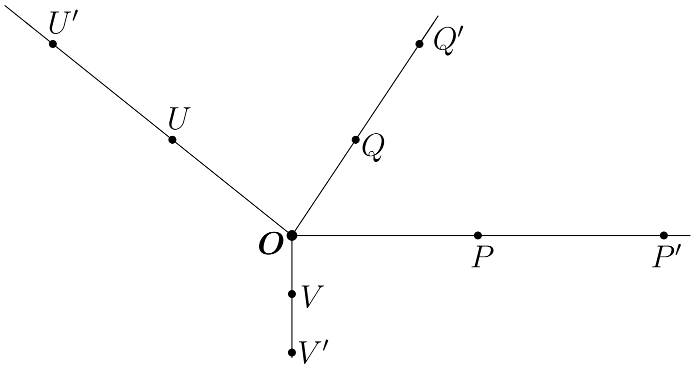
4 Dilation: Part I

The last section deals with isometries, but in this section, we will discuss motions that move points in the plane in order to increase or decrease distance without, in some sense, distorting shape. For rectilinear figures like triangles, size modification can be fully characterized by the ratios of corresponding sides, For a *curved* geometric figure like an ellipse, what it means for one ellipse to be "twice the size" of another one is less clear. The key concept involved is a *dilation.* A motion, or a rule, D of the plane is **a dilation with center O and scale factor r ($r > 0$)** if

(i) D leaves the point O fixed (i.e., $D(O) = O$).

(ii) If a point P is different from O, D moves P to the point P' on the ray R_{OP} so that $|OP'| = r|OP|$.

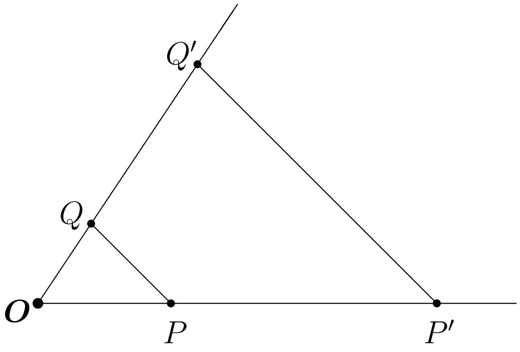
The standard notation for P' is $D(P)$. Thus a dilation with center at O maps each point by "pushing out" or "pulling in" the point along the ray from O to that point, depending on whether the scale factor r is bigger than 1 or smaller than 1. In particular, each ray issuing from O is mapped onto itself. Here is an example of how a dilation with $r = 2$ maps four different points (for any point P, we continue to let the corresponding letter with a prime, P', denote the image $D(P)$ of P):



The fundamental fact about dilations is the following. For the statement, we have to introduce the notation that $|\boldsymbol{AB}|$ stands for the **length** of the segment AB.

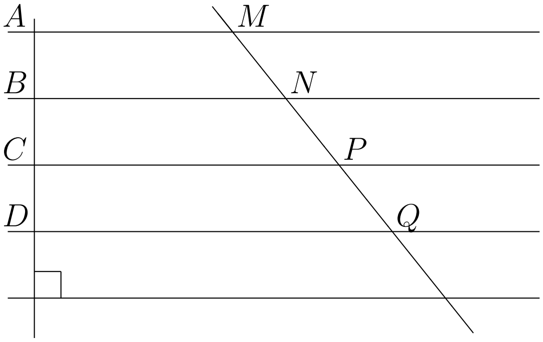
> (A) *If D is a dilation with center O and scale factor r, then for any two points P, Q in the plane so that O, P, Q are not collinear, the lines L_{PQ} and $L_{P'Q'}$ are parallel, where $P' = D(P)$ and $Q' = D(Q)$, and furthermore, $|P'Q'| = r\,|PQ|$.*

Activity 1 Check this statement experimentally by direct measurements for dilations with scale factors 2, 3, and something exotic like 3.7. (However, if you do it with 3.7, use a calculator!) Thus for the following example of a dilation with scale factor 3, one can use plastic triangles (see §2) to check that $PQ \parallel P'Q'$, and also check $|P'Q'| = 3|PQ|$.



One should appreciate why (A) is so remarkable: the definition of how a dilation moves the points in the plane involves only the "radial direction" with respect to the center of the dilation. In other words, if the center of the dilation D is O and P is a point distinct from O, then to find where $P'(= D(P))$ is, all you need is to look at O *and nothing else*: if you know where O is, you know how to draw the ray R_{OP} and on this ray you can get hold of P'. The same holds for Q'. What (A) says is that, although P' appears to have nothing to do with Q', yet the segment $P'Q'$ will always be tied to PQ in that $PQ \parallel P'Q'$ and $|P'Q'| = r\,|PQ|$, where r is the scale factor of D.

The above activity can be done with great accuracy for any scale factor. If we set our collective sights lower by not insisting on great accuracy and are happy with using only scale factors which are fractions with small whole numbers in the numerator and denominator, then your lined notebooks provide a fertile playground for dilation activities. Here is an explanation. The lines on your notebook papers are supposed to be mutually parallel and **equi-distant**, i.e., if you draw a line L_{AB} perpendicular to one (and therefore *every*) line, then the segments intercepted by the lines on L_{AD} are all of the same length. Thus $|AB| = |BC| = |CD| = \ldots$ (see the picture below). Now draw another line L_{MQ} at random but make sure that it intersects all the given lines and check (using a compass, for instance) that, always, we have $|MN| = |NP| = |PQ| = \ldots$. Such a line as L_{MQ} is called a **transversal** of the parallel lines.



Our finding may therefore be re-phrased as follows:

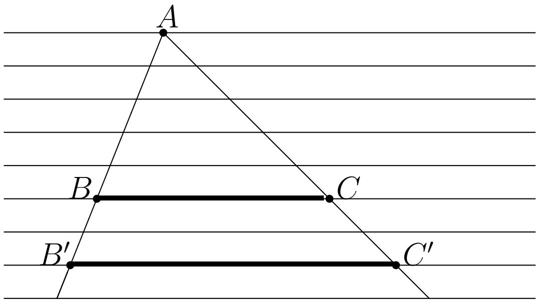
(B) *Equi-distant parallel lines intercept equal (length) line segments on a transversal.*

In a problem of Exercises 5.3 in the next chapter, you will have a chance to prove (B). In any case, we are now in a position to describe an activity on how to use your lined notebooks.

Activity 2 On a lined notebook paper, take a point A on a line and draw two rays R_{AB} and R_{AC} from A. Let both B and C be on (let us say) the 5th line below A. Let the intersections of the rays R_{AB} and R_{AC} with the 7th line below A be B' and C', respectively (see picture below). Then according to (B), we have

$$\frac{|AB'|}{|AB|} = \frac{|AC'|}{|AC|} \quad \text{because both are equal to } \frac{7}{5}.$$

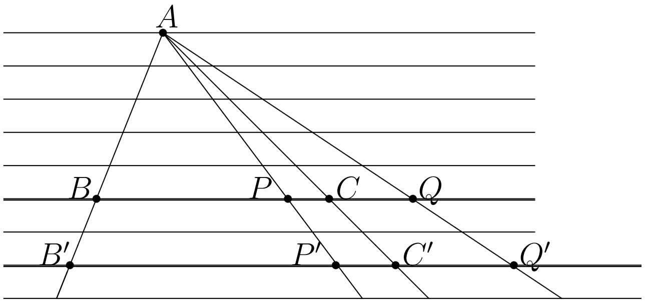
Thus if D^* is the dilation with center A and scale factor $\frac{7}{5}$, then $D^*(B) = B'$ and $D^*(C) = C'$. Now check by direct measurements that $|B'C'| = \frac{7}{5}|BC|$. Repeat this activity by varying the numbers 5 and 7.



The reason we said this activity with notebook papers may not be of great accuracy is that one cannot always be sure that the lines are truly equi-distant.

The next striking fact about dilations is somewhat subtle. Let us go over the above statement (A) with care: We have dilation D and two points P and Q so that the center O is not collinear with P and Q. We get two more points $P' = D(P)$ and $Q' = D(Q)$. Then (A) says the lines L_{PQ} and $L_{P'Q'}$ are parallel, *but it says nothing about the image under D of PQ*. Recall that the point $D(V)$, for any point V, is called the **image of V under D**, and that by the **image of PQ under D**, we mean the collection of the images of all the points in the line segment PQ under D. The notation for the latter is, as usual, $\boldsymbol{D(PQ)}$. Thus far, we have looked into the segment $P'Q'$ joining P' $(= D(P))$ to Q' $(= D(Q))$, but we have no idea what $D(PQ)$ might be. In other words, if $A \in PQ$, do we know that $D(A)$ lies in the segment $P'Q'$? You should try to find out:

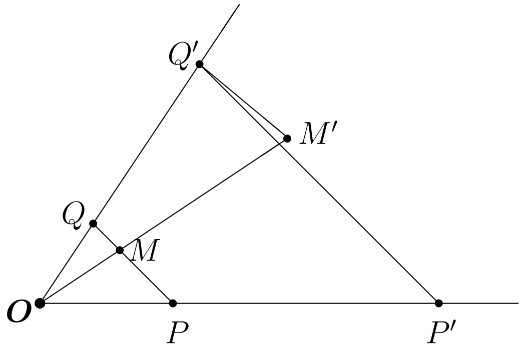
Activity 3 We pursue the example of Activity 2. Pick any point P on L_{BC}, and let the line joining A and P intersect $L_{P'Q'}$ at P'. Now measure $|AP|$ and $|AP'|$; is it true that $|AP'| = \frac{7}{5}|AP|$? Pick another point Q on L_{PQ} and get Q' on $L_{P'Q'}$ as shown. Again, is it true that $|AQ'| = \frac{7}{5}|AQ|$? Try other choices P and Q.



What turns out to be true in general is that if D is any dilation and P, Q are any two points, then

$$D(L_{PQ}) \;=\; L_{P'Q'} \quad \text{where } P' = D(P) \text{ and } Q' = D(Q). \tag{8}$$

We give the simple proof right now, assuming the truth of (A), as follows. Take any point M on L_{PQ}, and let M' be the image under D of M. We want to show that M' lies on $L_{P'Q'}$, where $P' = D(P)$ and $Q' = D(Q)$.



Because $Q' = D(Q)$ and $M' = D(M)$, (A) implies that $L_{Q'M'} \parallel L_{QM}$. Because $Q' = D(Q)$ and $P' = D(P)$, (A) implies that $L_{Q'P'} \parallel L_{QP}$. Of course $L_{QM} = L_{QP}$, we now have two lines $L_{Q'M'}$ and $L_{Q'P'}$ passing through the point Q' and both parallel to the same line L_{QP}. We know that *there can be only one line passing through a point and parallel to a given line.* So $L_{Q'M'} = L_{Q'P'}$ and M' lies on $L_{P'Q'}$. Therefore for any $M \in L_{PQ}$, we have proved that $D(M) \in L_{P'Q'}$. In other words,

$$D(L_{PQ}) \ \subset \ L_{P'Q'} \quad \text{where } P' = D(P) \text{ and } Q' = D(Q).$$

We leave the proof of the converse statement that $L_{P'Q'} \subset D(L_{PQ})$ to an exercise. We have thus proved equation (8).

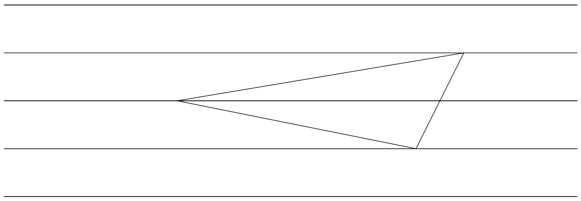
With a little bit more work, we can prove that, in fact, the image of the *segment* PQ under D is the *segment* $P'Q'$. There is an analogous statement about rays.

We summarize this discussion in the following statement:

> (C) *The image under a dilation of a line is a line, and the image of a segment is a segment and of a ray is a ray.*

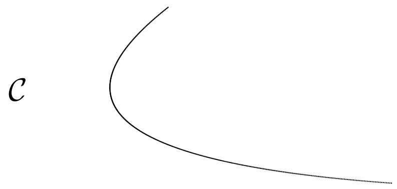
Statement (C) makes it extremely easy to find the image of a triangle under a dilation: just find the images of the three vertices (repeat: *only* three points) and then connect them to get the image triangle.

Activity 4 (1) Copy the following triangle into your lined notebook. Choosing any point *on a line* as the center of dilation, dilate the triangle with a scale factor of 3. Does it look *similar* to the original triangle?



(2) Still with scale factor 3, repeat part (1) using another center of dilation (but make sure that it still lies on a line of your notebook paper). How do the two triangles compare?

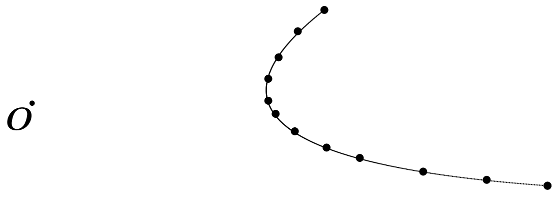
From Activity 4, you should get a sense that the dilated image of a geometric figure looks like the original figure and that the "shape" of the dilated image is independent of the center of dilation if the scale factor remains the same. We will see in Chapter 6 that dilation lies at the heart of the concept of *similarity.* This then brings us back to the point raised at the beginning of this section: if we have a *curved* figure, such as the curve $\mathcal{C}$ below, how to magnify it to be "twice as big"?



The answer is: dilate it (using any point as center of dilation) with a scale factor of 2. And of course, if you want to magnify $\mathcal{C}$ to make it 14.7 times bigger, then you dilate it with a scale factor of 14.7. ***Dilation is the method used to magnify or shrink geometric figures,*** using a scale factor > 1 for magnification, of course, and a scale factor < 1 for shrinking.

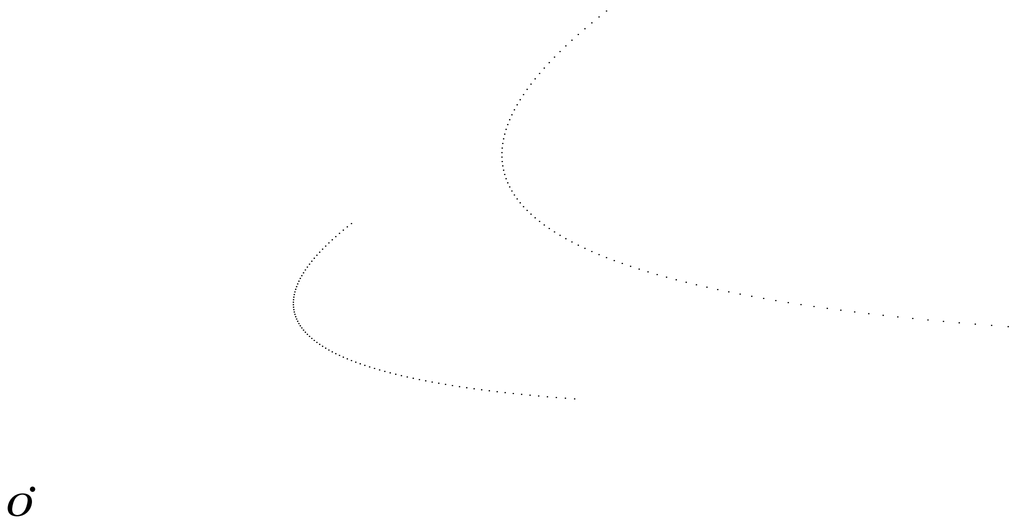
If D is a dilation with scale factor 2, then $D(\mathcal{C})$ is the collection of *all* the dilated images of the points on $\mathcal{C}$. Because $\mathcal{C}$ has an infinite number of points, it is not possible to draw $D(\mathcal{C})$, literally. In practice, we just draw the dilated images of *enough* points on $\mathcal{C}$ to get an idea of what $D(\mathcal{C})$ is like. With this in mind, we do the following:

Activity 5 Choose 12 points on $\mathcal{C}$ as shown, and choose some point O as center of dilation:

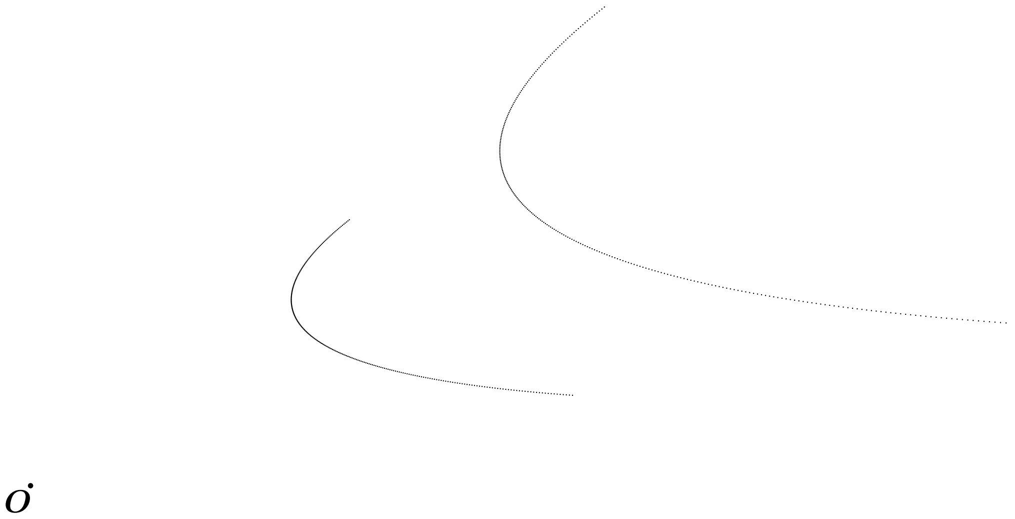


Now copy this picture on a piece of paper or transparency and dilate these 12 points with a scale factor of 2. Can you see the general shape of $D(\mathcal{C})$ just from these 12 dilated image points?

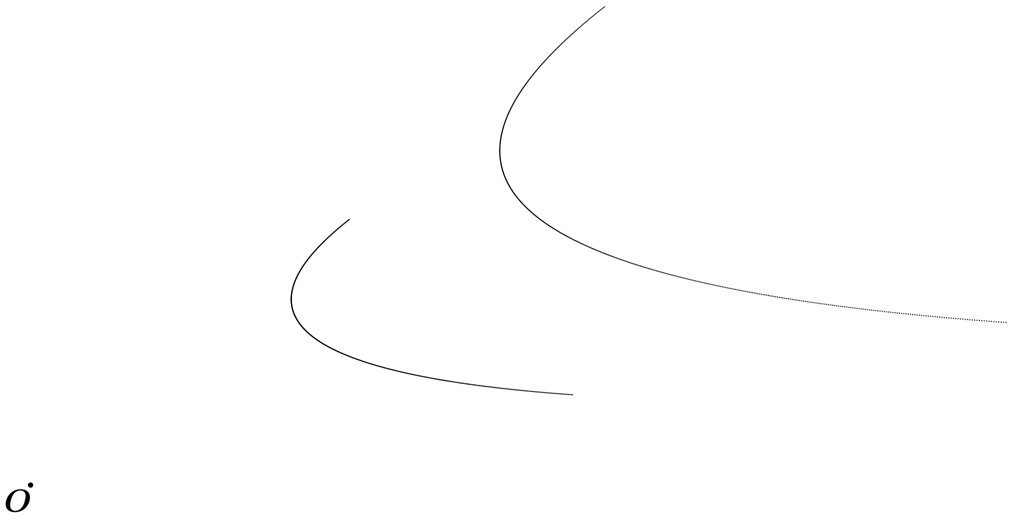
Because it is impractical to dilate too many points by hand, we are going to do a few elaborate magnifications by using the computer in order to impress on you the efficacy of dilating only a finite number of **data points** (the name we give to the chosen points on the original geometric figure) in a curved figure. It is obvious that the more data points we use, the better we can approximate the dilated curve by the image points. With the computer at work, it would be silly to use a simple scale factor of 2, so let us do something fancy by *dilating it with a scale factor of* 1.8. We will start with a modest number of 90 data points (there is a software reason for using 90) and dilate them from O, as shown. (We omit the rays joining each data point to O in the interest of visual clarity.) We will gradually increase the number of data points so you will get an idea of this process:



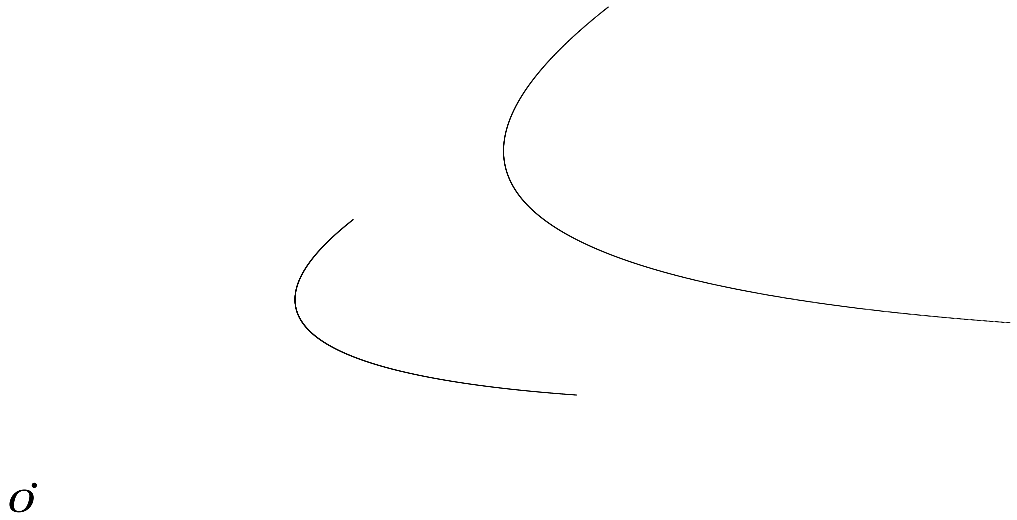
Next we triple the number of data points and use 270 instead of 90. The resulting approximation by the images of this finite collection of points to the image curve itself is already remarkably good.



If we use 600 points, then the images can almost pass for the real thing except that, if we look very carefully, we can still see discrete dots rather than a smooth piece of curve near the tail end of the longer curve.



Finally, if we use 1200 data points, then to the naked eye, these are two smooth curves, one being the dilation of the other. For all practical purposes, this approximation to the true dilated curve *is* the real thing.

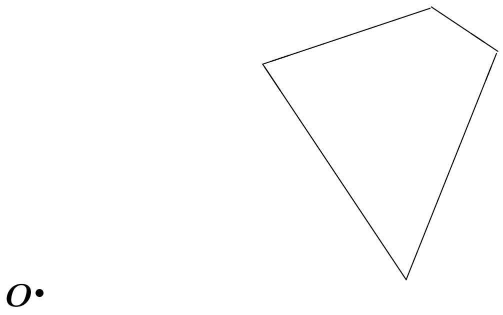


What we have described is a basic principle of constructing the dilated image of any object: To dilate a given object by a scale factor of r, replace the object by a *finite* collection of judiciously chosen data points, and then simply dilate these data points one by one by a scale factor of r. By increasing the number of data points, their dilated images yield a closer and closer approximation to the true dilated object. *This is how we can draw similar figures regardless of how curved they may be.* This is also the basic operating principle behind digital photography: approximate any real object by a large number of data points on the object, and then magnify or shrink these data points by dilation.

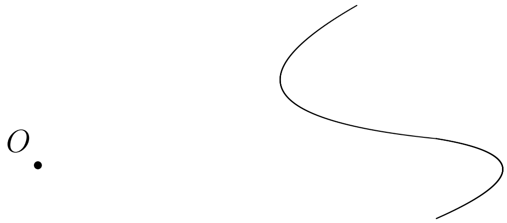
It is very instructive for school students to learn to magnify or shrink simple curved figures by such hands-on activities. These activities will not only impress them but also give them a far better conception of what "similarity" means than "same shape but not same size".

Exercises 4.4

1. Copy the following picture on a piece of paper and dilate the quadrilateral from O by a scale factor of $\frac{2}{3}$ (use a calculator):

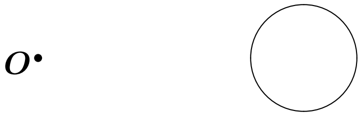


2. Given a point O and the following curve in the plane:



(a) Trace both on a piece of paper, and choose 10 points on the curve so that, by dilating these points with center O and scale factor 2, the dilated points give a reasonable picture of the dilated curve with scale factor 2. (b) Repeat part (a) by using 20 points. (You will notice that if you use only 10 data points, the points will have to be placed strategically in order that their images combine to give a good idea of the image curve.)

3. Dilate the following circle from O with a scale factor of 2.3. What is the dilated figure, and *why*? (*Caution:* This is a much harder problem than meets the eye.)



4. Complete the proof of equation (8) by proving $L_{P'Q'} \subset D(L_{PQ})$. In other words, if M' is a point on $L_{P'Q'}$, then there is some point $M \in L_{PQ}$ so that $D(M) = M'$. (*Hint:* Let M be the intersection of L_{PQ} and the ray $R_{OM'}$, and imitate the first part of the proof.)

5. Explain in detail the statement made below statement (C) to the effect that, to find the image of a triangle under a dilation, it suffices to find the images of the three vertices and then connect them to get the image triangle.

6. If you remember any high school geometry, try to explain the phenomenon in part (2) of Activity 4.

Chapter 5: Basic Isometries and Congruence

In this chapter, we begin the formal study of the geometry of the plane. The goal is to achieve at least a working knowledge of the meaning of *congruence*. The next chapter will tackle the related concept of *similarity*.

School geometry is the analytic and symbolic study of our visual perception of the space around us. In order to faithfully capture spatial information, we need to use very precise language because our reasoning has to be conducted using this verbal description of space.[41] We will give new definitions to many concepts already familiar to you, such as "half-planes", "angles", "convex sets", "rectangles", "polygons", etc. There is an inherent danger here that, because these terms *seem* so familiar, you take the new definitions for granted and ignore them. ***You'd better not***, because in an overwhelming majority of the cases, the new definitions are more precise than the ones you already know. Please pay special attention to the higher level of precision. To give an example, you may know a rectangle as "a quadrilateral with four right angles and two pairs of equal opposite sides", but here there will be two surprises. One, a rectangle is merely "a quadrilateral with four right angles" but there is no mention about equal sides, and two, you are going to get a dire warning that *a priori* we have no idea whether there are any rectangles or not. Eventually we will prove that rectangles do exist and that, indeed, they have equal opposite sides. *Please be aware of these new features when we begin this tour of geometry.*

In standard middle school textbooks, congruence is nothing but *same size and same shape*. To the extent that this is not a mathematically acceptable definition, we are forced to conclude that most middle school students do not have a clear idea of what congruence is. In high school geometry courses, congruence tends to become synonymous with "triangle congruence", and such a turn of events again leaves unexplained the general concept of congruence between curved figures.

In this chapter, we define congruence in terms of the basic isometries — rotations, translations, and reflections — in the plane. We begin with a precise definition of these three concepts. Since we have already devoted a good deal of effort in the last chapter to an intuitive, hands-on discussion of these isometries, it is hoped that the precise definitions, while somewhat sophisticated, can be taken for what they are: a

[41] In a strict sense, this is correct. But because a literal adherence to this ideal would make the teaching of geometry in schools painful if not impossible, we try to walk a thin line between what is correct and what is learnable. Compare H. Wu, *What is mathematics education?* http://math.berkeley.edu/∼wu/C49.pdf

serious attempt to capture in precise mathematical language the intuitive content of these motions of the plane. We need to explain congruence only because we need to explain similarity, and the reason for the latter is that we need some knowledge of similar triangles to make sense of the discussion of the slope of a line in beginning algebra. For this reason, we adopt a strictly utilitarian attitude toward the treatment of geometry: we only do enough to facilitate the learning of algebra.

Having pointed out the importance of precise geometric definitions, we are obligated to also emphasize that *the goal of geometry is not to study the precise definitions per se, but to understand the visual information encoded in the definitions.* The subsequent discussions will amply bear this out.

Unlike the first two chapters and Chapter 4, which present school mathematics at a level quite close to what can be taught in the school classroom, the material of the next three chapters does not as yet have a natural niche in the school curriculum. There is no doubt that most, if not all, of the reasoning is essential knowledge to a middle school teacher, but how much of it should be used in the middle school classroom has yet to be decided. The teaching of geometry in schools has been in disarray for quite some time and it is inadequate in terms of the proliferation of incorrect definitions and the lack of mathematical reasoning. What the remaining three chapters try to do is to fill in a little bit of this immense void and present geometry as a mathematical subject rather than as a glossary of mathematical terms or as a random collection of facts to be memorized. We hope you will make an effort to bring this viewpoint back to your classroom.

1 The basic vocabulary

The discussion of this section takes place in a fixed plane.

The basic assumptions

We are going to start from the beginning. In particular, we will give formal definitions of all the concepts we are going to use. In our subsequent mathematical explanations, it is understood that we draw *only* on these definitions if any of these concepts come up in the discussions.

By a **line**, we mean a *straight* line. We assume you know what a straight line is, but will nevertheless explicitly point out that it is understood to be *infinite in both directions*. We will be discussing lines lying in a fixed plane, and it would be a good idea if we have a common starting point. In the informal discussion of geometry in Chapter 4, for example, we took for granted the following facts, among many others:

> There is only one line joining two distinct points. (End of §3.)
>
> Through a point P passes only one line which is parallel to another line which does not contain P. (Proof of (C) in §4.

It stands to reason that, for the sake of transparency if nothing else, we put everything on the table and say exactly what we are going to take for granted in the forthcoming discussion. So here goes. There are six such statements, (L1)–(L6). You will undoubtedly agree that every single one of them is perfectly obvious.

(L1) *Through two distinct points passes a unique line.*

If L is the line passing through the points A and B, we say L **joins** A **to** B, and if there is any fear of confusion, we will write $\boldsymbol{L_{AB}}$ for $\boldsymbol{L}$. It follows from (L1) that:

(L2) *Two distinct lines intersect either at one point or none at all.*

This last statement is not completely satisfactory because it needs to be supplemented by a more precise statement about when two lines will intersect and when they will not. With this in mind, this is the right place to introduce one of the key definitions of plane geometry.

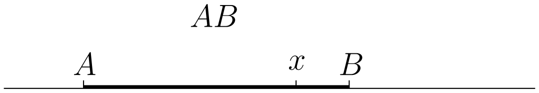
Definition *Two lines in the plane are said to be* **parallel** *if they do not intersect.*

In symbols, if two lines L and L' are parallel, we write $\boldsymbol{L \parallel L'}$. The following statement now completely clarifies the situation. It turns out to have profound implications in geometry as well as in the development of mathematics as a whole.

(L3) (**Parallel Postulate**) *Given a line L and a point P* not *on L but lying in the same plane, there is exactly one line in the plane passing through P which is parallel to L.*

In other words, we assume as obvious that in the plane that we normally work with, for a point P not on a line L, every line containing P intersects L except for one line. You will see that the Parallel Postulate dominates the discussion of plane geometry.

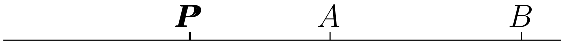
If A and B are points on a line L, denote by $\boldsymbol{AB}$ the collection of all the points on L between A and B, together with the points A and B themselves. We call AB the **line segment**, or more simply the **segment** joining A and B, and the points A and B are called the **endpoints** of the segment AB. Note that it makes sense to talk about points on L *between* A and B, because L may be regarded as a number line and A and B then become numbers, e.g., $A = 0$ and $B > 0$. In that case, the points between A and B would be all the numbers x so that $0 < x < B$.



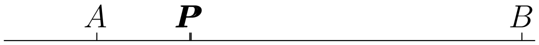
The next property of a line is most likely not one you have encountered in mathematical discussions thus far, possibly because it is deemed too obvious to be mentioned.

(L4) (**Line separation**) *A point P on a line L separates L into two* non-empty *subsets L^+ and L^-, called* **half-lines***, so that*

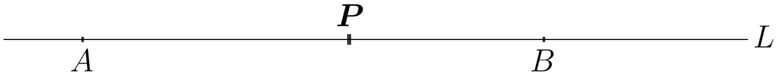
(*i*) *The line L is the* **disjoint union** *of L^+, $\{P\}$, and L^-, in the sense that every point of L is in one and only one of these three sets* (*in particular, the three sets L^+, $\{P\}$, and L^- are disjoint*).
(*ii*) *If two points A, B belong to the same half-line, then the line segment AB does not contain P,*



(*iii*) *If two points A and B belong to different half-lines, then the line segment AB contains P.*



This property of the line is best understood from the perspective that every line is a number line. Thus if we have a point P on a line L , as shown, then letting P be 0, we may let L^+ be all the positive numbers (those > 0), and L^- be all the negative numbers (those < 0). In this case, if $A \in L^-$ and $B \in L^+$, then one sees that the segment AB must contain P $(= 0)$.



Of course, if A and B are both positive or both negative numbers (so that $A, B \in L^+$ or $A, B \in L^-$), AB would not contain P.

The preceding identification of a line with a number line is of course the reason for the notation L^+ and L^-. One should not get the idea that, as a consequence of this fact, one should always think of L^- as the half-line on the left and L^+ as the half-line on the right. This only works if the line is “horizontal” (and we have yet to make formal sense of this terminology). For a line that is “vertical” or almost so, it would be impossible to carry out this analogy.

The union of either half-line, L^+ or L^-, with P is called a **ray**. We also say these are rays **issuing from $\boldsymbol{P}$**. If we want to specifically refer to the ray containing A, we use the symbol $\boldsymbol{R_{PA}}$. Similarly, the ray containing $\boldsymbol{B}$ issuing from $\boldsymbol{P}$ is denoted by

$\boldsymbol{R_{PB}}$. The point P is the **vertex** of either ray. The two rays have only the point P in common, and each ray is infinite in only one direction.

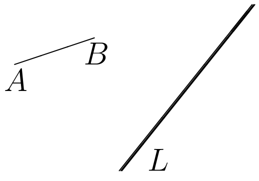
If a point P lies on a line L and A, B are any two points on L not equal to P, then (L4) tells us how to determine whether two distinct points A and B are on opposite half-lines of L relative to P: they are on opposite half-lines if and only if AB contains P, and are in the same half-line if and only if AB does not contain P. This observation will be used frequently below.

You may wonder why, if a point P on a line L is given, we don't just identify L with a number line and P with 0 so that, once that is done, we won't even have to bother mentioning something as obvious as (L4). One reason is that if each time we see a line, we have to make clear what the identification is (i.e., which direction is positive), it gets tedious. And if we consider several lines all at once? Then you begin to worry about how to make the identifications in a "nice" consistent way, which then becomes a nonessential distraction. So what (L4) does is to provide a direct description of the separation of a line by a point lying in it *independent of any identification with a number line.* The same idea lies behind the next property of a line, which describes the relationship between a line and the plane containing it.

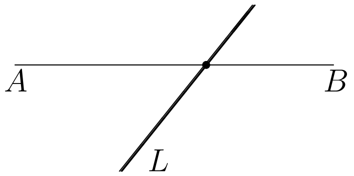
(L5) (**Plane Separation**) *A line L divides the plane into two* non-empty *subsets, $\mathcal{L}$ and $\mathcal{R}$, called* **half-planes**. *These half-planes have the following properties*:

(*i*) *The plane is the* **disjoint union** *of L, $\mathcal{L}$, and $\mathcal{R}$,* i.e., the union of L, $\mathcal{L}$, and $\mathcal{R}$ is the whole plane and no two of these sets have any point in common.

(*ii*) *If two points A and B in the plane belong to the same half-plane, then the line segment AB does not intersect L.*



(*iii*) *If two points A and B in the plane belong to different half-planes, then the line segment AB must intersect the line L.*



A comment about (L5) is in order. Clearly, one would prefer a more explicit description of the half-planes of a line. After all, if a line is drawn on a piece of paper or on a blackboard, one can point to the two "halves" of the plane separated by the line. In a *middle school* classroom, this is what you should do without a doubt: just point to the half-planes and not burden these students with abstract statements like (L5). But as a teacher, you should learn to appreciate the difficulty of transcribing the visual information into precise and (in this case) abstract language. Without waving our hands about what is "on the left" or what is "on the right", we learn to use properties (*i*)–(*iii*) above to pin down *precisely* what these half-planes are.[42] Although they are non-intuitive, (*i*)–(*iii*) nevertheless leave no doubt that each half-plane is exactly what our intuition says it is: all the points "on one side of L". More precisely, suppose we are given two points A and B in the plane, neither lying on L and they "lie on opposite sides of L" in the sense that the segment AB intersects L. Then according to (*ii*), A and B lie in different half-planes. Moreover, (*iii*) tells us that if two points C and D are in the same half-plane, then they do not "lie on opposite sides of L" in the above intuitive sense. Without more information about a line, this is all we can do about its half-planes. However, once we have coordinates and we can describe a line by an equation, then we will be able to describe the half-planes of a line explicitly. This will be a topic for a future discussion.

The union of either $\mathcal{L}$ or $\mathcal{R}$ with L is called a **closed half-plane**.

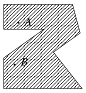
In the normal way of drawing a plane, if the line L goes left-right, then we will continue to refer to it informally as *horizontal*, and the half-planes are referred to as the *upper half-plane* and the *lower half-plane*. If L goes up-down, then it will continue to be referred to as a *vertical* line and the half-planes are the *left half-plane*

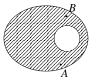
[42] We will use the same abstract idea once more at the end of this section for the statement of the Theorem about polygons.

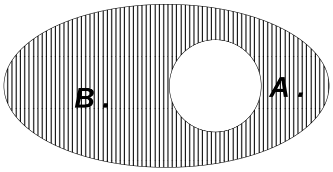
and the *right half-plane*. Again, all this can be made more precise in the presence of a coordinate system.

Angles and the crossbar axiom

The last property of lines that we assume to be known cannot be stated until we have the (always troublesome) concept of an angle. To this end, we first introduce the concept of convexity. A subset $\mathcal{R}$ in a plane is called **convex** if given any two points A, B in $\mathcal{R}$, the segment AB lies completely in $\mathcal{R}$.[43] The definition has the obvious advantage of being simple to use, so the concern with this definition is whether or not it captures the intuitive feeling of "convexity". Through applications, you will see that it does. Every line and the plane itself are of course convex. The half-lines and half-planes are convex, by virtue of the properties (L4) and (L5) above, respectively (exercise). It is also an elementary exercise to show that each closed half-plane is convex. Many common figures, such as the interior of a triangle or a rectangle or a circle, once they have been properly defined, will also be seen to be convex. The following shaded subsets of the plane are, however, not convex, because, visibly, the segment AB of the points A and B in each region does not lie entirely in the respective shaded subset itself.







The intersection of two half-planes or two closed half-planes is easily seen to be convex (see problem 2 in Exercises 5.1 for a more comprehensive statement). This fact, together with the concepts of rays and half-planes enter into the following definition of the concept of an angle.

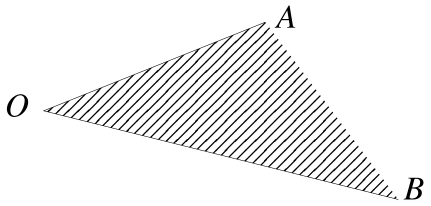
[43] This definition is in fact valid in any dimension.

Given three points O, A, B in the plane which are not **collinear**, i.e., do not lie on a line, let R_{OA}, R_{OB} be two rays issuing from O. These rays determine two subsets of the plane. One of them is the intersection of the following two closed half-planes:

the closed half-plane of the line L_{OA} containing B, and

the closed half-plane of the line L_{OB} containing A.

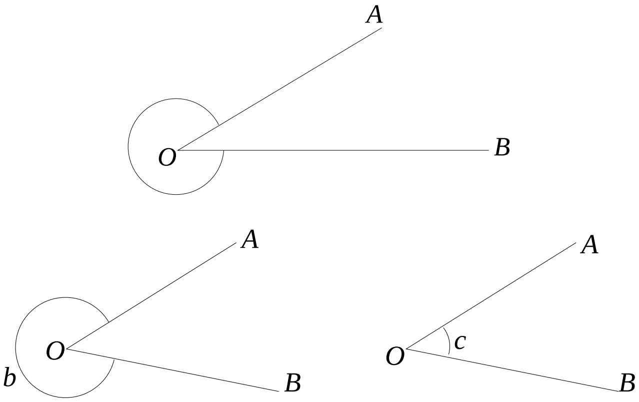
By the observation above, this is a convex set, and is suggested by the shaded set in the following figure.



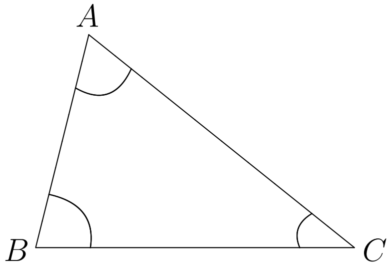
The other subset determined by R_{OA} and R_{OB} is the union of the **complement**[44] of the shaded set together with the two rays R_{OA} and R_{OB}. This is suggested by the unshaded set in the above figure, which is visibly not convex . Then either the convex or the nonconvex subset determined by these two rays R_{OA} and R_{OB} is called the **angle determined by these rays.** These rays are called the **sides** of the angles, and the point O is the **vertex** of the angle. We emphasize that, in this book, an angle is always one of the subsets of the plane "between" the two rays rather than just the union of the two rays themselves. *Unless stated otherwise, we follow the standard practice of taking the convex subset* (*the shaded subset*) *to be the angle* and denote it by $\angle \boldsymbol{AOB}$. If we want to consider the angle determined by the non-convex subset, we would have to say so explicitly or use an arc to so indicate, e.g.,

A better notation, one that will be used often, is to use an arc *and* a letter in the region to indicate the angle. Thus $\angle b$ denotes the nonconvex region below, while $\angle c$ denotes the convex region.

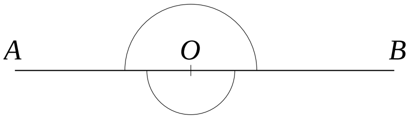
[44] The **complement** of a set S in the plane is by definition the set of all the points in the plane not lying in S.



Assuming you know what a *triangle* is (a precise definition will be given presently), this way of denoting angles is especially relevant in the case of a triangle ABC. In this case, since the angles $\angle ABC$, $\angle ACB$, and $\angle BAC$ are understood to be convex, each must contain triangle ABC itself, and we usually let $\angle \boldsymbol{A}$ to stand for $\angle BAC$, $\angle \boldsymbol{B}$ for $\angle ABC$, and $\angle \boldsymbol{C}$ for $\angle ACB$.

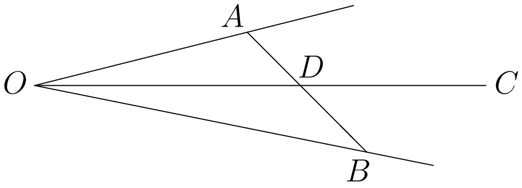


If O, A, B are collinear, then either A and B are in the same half-line with respect to O, or in opposite half-lines. In the former case, we call it a **zero angle**. In the latter case, either half-plane determined by the line containing these points will be called a **straight angle**.



The last assumption on the properties of lines will now be stated:

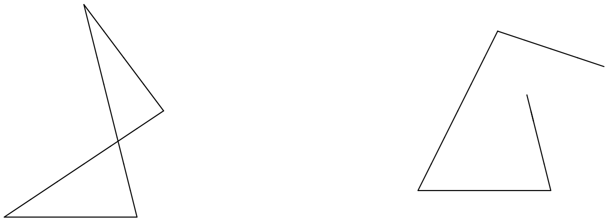
(L6) (**Crossbar axiom**) *Given angle AOB, then for any point C in $\angle AOB$, the ray R_{OC} intersects the segment AB* (indicated by the point D in the following figure).



You may regard the crossbar axiom as frivolous, because "what else can the ray R_{OC} do"? First of all, so long as you consider this statement to be obvious, our objective of agreeing on a common starting point will be met. As to whether the crossbar axiom is frivolous, we should point out that up to this point, none of (L1)–(L5) guarantees that the ray R_{OC} must intersect AB. It is the purpose of the crossbar axiom to make official the intuitive idea that a ray is indeed "straight" and therefore must meet a segment that is "in front of it". For example, if we assume for a moment we know what the angle bisector of an angle is (a concept that will be defined presently), then (L6) guarantees that the angle bisector of an angle in a triangle must intersect the opposite side. This is reassuring.[45]

Polygons

The next basic concept we need is that of a polygon. Intuitively, we do not want the following figure on the left to be called a "polygon" because it "crosses itself", and we do not want the following figure on the right to be a "polygon" either because it "doesn't close up".

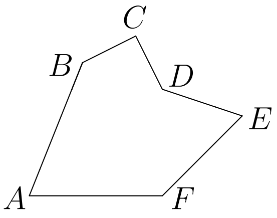


It is clear that the definition of a polygon requires some care. We first define a special case of a polygon: a **hexagon** is by definition a collection of six points A, B, C, D, E, F in the plane together with the six segments

[45] We note that in a strictly logical development of plane Euclidean geometry, the crossbar axiom can be deduced from the plane separation property (L5). See page 116 of M. J. Greenberg, *Euclidean and Non-Euclidean Geometry*, 4th Edition, W. H. Freeman, 2008. However the proof is too technical to be of any educational value to most prospective teachers.

$$AB,\ BC,\ CD,\ DE,\ EF, \text{ and } FA,$$

so that *any two of them do not intersect each other except at the endpoints as indicated,* i.e., CD intersects DE at D, DE intersects EF at E, etc.



The six points A, B, $\ldots$, F, are called the **vertices** of the hexagon and the six segments AB, BC, $\ldots$, FA are its **sides** or **edges**. Notice that by its very definition, a hexagon labels its vertices **cyclically** in the sense that its sides connect all of them in alphabetical order until the very end, when the last vertex F is connected to the first vertex A.

Now that we have defined a hexagon, we want to define a polygon of any number of sides (or vertices, for that matter). Then we come up against a problem with *notation*: for six vertices, we can employ A, $\ldots$, F, but if we have a polygon with 234 vertices, what symbols should we employ to denote these vertices? We can use numbers instead of letters to denote the vertices, in which case, we can go from 1, 2, …all the way to 234. But because integers come up in so many contexts, sooner or later this would lead to hopeless notational confusion. We are therefore forced into using subscripts: we can efficiently denote the 234 vertices by the 234 symbols A_1, A_2, A_3, $\ldots$, A_{233}, A_{234}. Of course we could have used any letter, say V, instead of A for this purpose, e.g., V_1, V_2, V_3, $\ldots$, V_{233}, V_{234}.

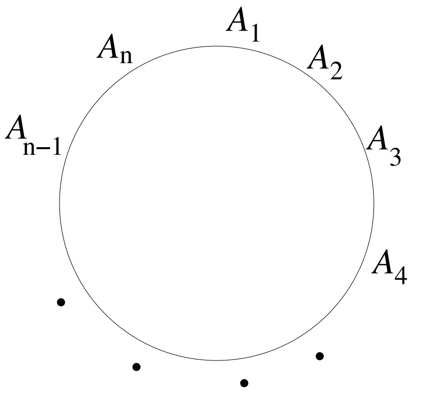
We can now give the general definition of a polygon and related concepts.

Let n be any positive integer ≥ 3. An **n-sided polygon** (or more simply an **n-gon**) is a collection of n distinct points A_1, A_2, $\ldots$, A_n *in the plane*, together with the n segments A_1A_2, A_2A_3, $\ldots$, $A_{n-1}A_n$, A_nA_1, so that *none of these segments intersects any other except at the endpoints as indicated*, i.e., A_1A_2 intersects A_2A_3 at A_2, A_2A_3 intersects A_3A_4 at A_3, etc. The n-gon is denoted by $\boldsymbol{A_1A_2\cdots A_n}$. If

$n = 3$, the polygon is called a **triangle**; $n = 4$, a **quadrilateral**; $n = 5$, a **pentagon**; and if $n = 6$, a **hexagon**, as we have seen. These names came to us from Euclid's **Elements**, and in principle there is a name for every n-gon. For example, if $n = 10$, the polygon is called a **decagon**. But such extra erudition is not a necessity since **10-gon** would do just fine. So unless absolutely necessary, we will only use the Greek names of the first six polygons.

Given polygon $A_1 A_2 \cdots A_n$, as in the earlier case of the hexagon, the A_i's are called the **vertices** and the segments A_1A_2, A_2A_3, etc. the **edges** or sometimes the **sides**. For each A_i, both A_{i-1} and A_{i+1} are called its **adjacent vertices** (except that in the case of A_1, its adjacent vertices are A_n and A_2, and in the case of A_n, its adjacent vertices are A_1 and A_{n-1}). Thus the sides of a polygon are exactly the segments joining adjacent vertices. Any line segment joining two nonadjacent vertices is called a **diagonal**.

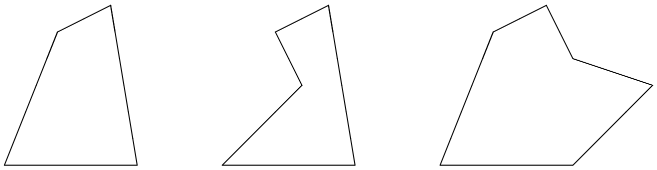
The best way to remember the notation associated with a polygon is to think of the points A_1, A_2, ..., A_n as being placed in sequence around a circle,[46] for example, in clockwise (or counterclockwise) direction:



[46] Note that, here, we are using the concept of a "circle" in an informal way. The formal definition will be given later.

Then it is quite clear from this arrangement whether or not two vertices are adjacent.

The following are examples of polygons (with the labeling of the A_i's omitted):



Geometric measurements

We next address the issue of measurement: how to measure the length of segments and degrees of angles.

We begin with length. When we dealt with the number line, we could choose any segment to be the unit segment, i.e., we could declare any length to be 1. That is because if we only deal with one line, such a decision affects only what is done on that line. Now that we have to deal with the plane which has many lines, the choice of a segment of unit length on one line will have to be consistent with the choices on other lines in order to make possible the discussion of length *in the plane*.

> **Assume that we can decree one choice of a unit segment, once and for all, on *all* the lines, and consider this done.**

Then every line can be considered to be the number line and every segment AB on a line L now has an unambiguous **length**, to be denoted by $|\boldsymbol{AB}|$. This has many implications. The first one is that all that we have learned about the number line can now be transferred to each line in the plane. For example, let A and B be points on a line L, and suppose $|AB| = r$ for some positive number r. If we consider L as the number line, and take A to be 0 with B in the positive direction, then the segment AB coincides exactly with the segment $[0, r]$. Note that this also implies there is another point on L, to be called B', so that B and B' are on different rays issuing from A and $|AB| = |AB'|$. Indeed, simply take B' to be the point corresponding to $-r$ when L is regarded as the number line as above.[47]

[47] To anticipate a future development, the ability to look at each line as the number line is one reason that we can set up coordinate axes in the plane.

Another consequence of the universal adoption of a unit length for all lines in the plane is the possibility of defining the concept of **distance** between any two points A and B in the plane, denoted by $\mathbf{dist}(\boldsymbol{A}, \boldsymbol{B})$: it is the length of the segment between A and B on the line L_{AB}. Clearly:

(D1) $\operatorname{dist}(A, B) = \operatorname{dist}(B, A)$.

(D2) $\operatorname{dist}(A, B) \geq 0$, $\;$ and $\operatorname{dist}(A, B) = 0 \iff A$ and B coincide.

(D3) If A, B, C are collinear points, and C is between A and B, then

$$\operatorname{dist}(A, B) = \operatorname{dist}(A, C) + \operatorname{dist}(C, B)$$

Note that (D2) implies that if we want to prove $A = B$, all we have to do is to prove that $\operatorname{dist}(A, B) = 0$. This may sound trivial, but it will turn out to be useful.

There are two more facts about distance that are equally basic but perhaps less obvious. We point them out here, but will not have time to give a proof.

(D4) (**Triangle inequality**) Any three points A, B, C satisfy the inequality $\operatorname{dist}(A, C) \leq \operatorname{dist}(A, B) + \operatorname{dist}(B, C)$.

(D5) For any three points A, B, C, the equality $\operatorname{dist}(A, C) = \operatorname{dist}(A, B) + \operatorname{dist}(B, C)$ holds if and only if A, B, C are collinear and B is between A and C.

With the availability of measurements for line segments, we can now formally introduce the concept of a circle. Fix a point O. Then the set of all points A in the plane so that $\operatorname{dist}(O, A)$ is a fixed positive constant r is called the **circle of radius $\boldsymbol{r}$** (in the plane) **about $\boldsymbol{O}$.** The point O is the **center**.

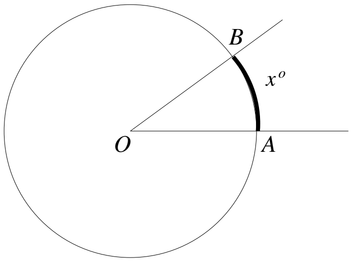
This is the official meaning of the concept of a circle, but in school mathematics, the word "circle" is usually used in an undisciplined way. It is used for both the circle as defined above *and* the **region enclosed by the circle**, which is defined precisely as

$$\text{all the points } A \text{ satisfying } \operatorname{dist}(A, O) \leq r.$$

In mathematics, the latter is called a **disk**, or **the disk with radius $\boldsymbol{r}$ and center $\boldsymbol{O}$**, to be precise. Thus when one talks about "the area of the circle", what is meant is in fact "the area of the disk". We usually go along with this kind of sloppiness when no harm is done, but if absolute clarity is mandatory, we will make the distinction between "circle" and "disk".

A circle whose radius is of length 1 is called a **unit circle**. Using a unit circle, we now describe how to measure the size of an angle by assigning it a degree. Divide the unit circle into 360 parts of equal length,[48] 360 **equal parts** for short. The length of one part is called **one degree**. Then we can subdivide a degree into n equal parts (where n is any whole number), thereby obtaining $\frac{1}{n}$ of a degree, etc. *It is exactly the same as the division of the chosen unit on a number line into fractions*, except that in this case, we have a "circular number line" and, once a point has been chosen to be 0, the number 360 coincides with 0 again. A single (connected) piece of a circle is called an **arc**.

Now suppose $\angle AOB$ is given, then it intercepts an arc on the unit circle around O (equivalently, this arc is the intersection of $\angle AOB$ with the unit circle around O).[49] For the sake of notational simplicity, let us assume that both A and B are points on the unit circle around O. In the picture below, B is in the counterclockwise direction from A, but if B happens to be in the clockwise direction from A, then the following discussion will have to be adjusted accordingly.



As in the case of the number line, we are free to choose A or B (or in fact any point on this unit circle) as the 0 of this "circular number line"; let us say A for definiteness. Because the arc intercepted by $\angle AOB$ on this unit circle is in the counterclockwise

[48] There is a subtle point in this definition. It has to do with the fact that "equal parts" here refer to "arcs of equal length", but we have yet to define the concepts of "arc" and "length of an arc". There is no fear of circular reasoning, however, because an arc and the length of an arc can be defined independently; see Chapter 7.

[49] At this point, you see the advantage of defining an angle as a region in the plane.

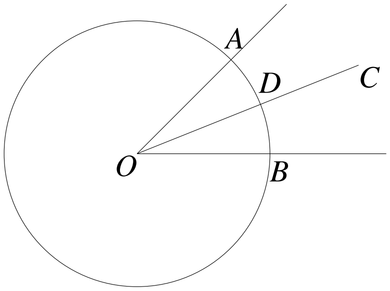
direction from A, we chose the point on the unit circle, which is 1 degree from A and in the counterclockwise direction from A, to be the unit 1 of this "circular number line". Then the degrees 1, 2, 3, ..., 359, 360 go around the unit circle in a clockwise direction until 360 comes back to 0 (which is A). Now, on this "circular number line", whose whole numbers up to 359 increase in the counterclockwise direction, B has a numerical value x. Then we say $\angle AOB$ **is** x **degrees** or that its **measure** or **magnitude** is x degrees, and we write $|\angle \boldsymbol{AOB}| = \boldsymbol{x}^\circ$, where the small circle $^\circ$ in the superscript position indicates that we are using degree as the unit. Thus by definition, *degree is a number* ≥ 0 *and* ≤ 360. For now, if we measure degrees in the counterclockwise direction (as we just did), we would explicitly say so, but if we measure in the clockwise direction we can do that too. In §2, we will adopt a uniform method of expressing whether we measure angles in the clockwise or counterclockwise direction. (There is another unit of measurement for measuring angles, called "radian", that is used almost universally in advanced mathematics.)

Notice that the method of angle measurement we have just described is exactly the principle used in the construction of the protractor.

The following useful facts about angles are usually deemed to be too obvious to be mentioned. We mention them explicitly because we will need them at crucial junctures and want it known that there is a reasoning behind them.

($\boldsymbol{\alpha}$) *If B and B' are two points lying in the same half-plane of L_{OA} and $|\angle AOB| = |\angle AOB'|$, then the rays R_{OB} and $R_{OB'}$ coincide.*

($\boldsymbol{\beta}$) *If C is a point in $\angle AOB$, then $|\angle AOC| + |\angle COB| = |\angle AOB|$.*

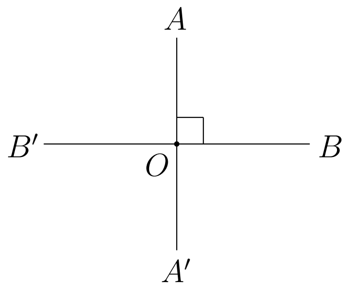


The reason for the first is that the rays R_{OB} and $R_{OB'}$ must intersect the unit circle around the point O at exactly the same point. The reason for the second is that if the ray R_{OC} intersects the unit circle around O at a point C', as shown in the figure above, then the degrees of the arcs (on the unit circle) from A to C' and from C' to B add up to the degree of the arc from A to B.

An angle is a zero angle exactly when it is 0°, and is a straight angle exactly when it is 180°. An angle of 90° is called a **right angle**. An angle is **acute** if it is less than 90°, and is **obtuse** if it is greater than 90°. There are analogs of these names for triangles, namely, a triangle is called a **right triangle** if one of its angles is a right angle, an **acute triangle** if all of its angles are acute, and an **obtuse triangle** if (at least) one of its angles is obtuse. (Anticipating the theorem that the sum of the degrees of the angles in a triangle is 180, we know that a triangle cannot have more than one obtuse angle or more than one right angle.)

We observe that our convention of taking every angle to be convex unless otherwise specified amounts to saying that, without statement to the contrary, an angle is at most 180°*.*

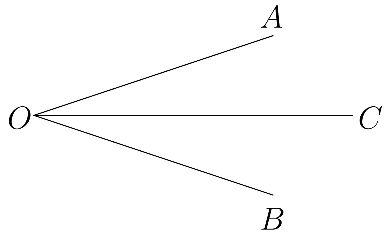
Let two lines meet at O, and suppose one of the four angles, say $\angle AOB$ as shown, is a right angle.



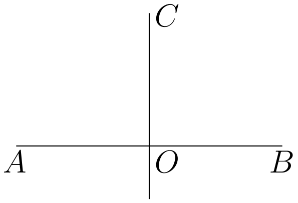
Because a straight angle is 180 degrees, it is easy to see that all the remaining angles are also right angles, i.e., $|\angle BOA'| = |\angle A'OB'| = |\angle B'OA| = 90°$. It follows that when two lines meet and one of the four angles so produced is a right angle, it is unambiguous to say that the two lines are **perpendicular**. In symbols: $\boldsymbol{L_{AO} \perp L_{OB}}$ in the notation of the preceding figure, although it is equally common to write instead, $\boldsymbol{AO \perp OB}$. A ray R_{OC} in an angle[50] AOB is called an **angle bisector** of $\angle AOB$

[50] Recall, an angle is a region.

if $|\angle AOC| = |\angle COB|$. Sometimes we also say less precisely that **the line $\boldsymbol{L_{OC}}$ (rather than the ray $\boldsymbol{R_{OC}}$) bisects the angle $\boldsymbol{AOB}$.**



It is clear that *an angle has one and only one angle bisector.* Therefore if $CO \perp AB$ as shown below, then CO is the unique angle bisector of the straight angle $\angle AOB$.



We thus have:

> *Let L be a line and O a point on L. Then there is one and only one line passing through O and perpendicular to L.*

With the availability of measurements for both angles and line segments, we can now complete the list of standard definitions.

If AB is a segment, then the point C in AB so that $|AC| = |CB|$ is called the **midpoint** of AB. Analogous to the angle bisector, the **perpendicular bisector** of a segment AB is the line perpendicular to L_{AB} and passing through the midpoint of AB. It follows from the uniqueness of the line perpendicular to a line passing through a given point that *there is one and only one perpendicular bisector of a segment.*

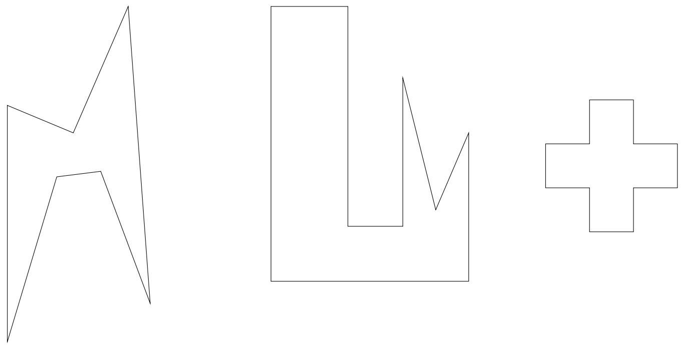
Next we turn to polygons. The **perimeter** of a polygon is the sum of the lengths of (all) its sides. This concept should be introduced as soon as the length of a segment is defined. There is a perception that students confuse the “perimeter” and “area” of a polygon, and one reason is undoubtedly the fact that these concepts are usually introduced together.[51] However the concept of area is more sophisticated (see Chapter 7), and we suggest that you decouple these two concepts in your classroom instruction.

We now introduce some common names for certain triangles and quadrilaterals. An **equilateral triangle** is a triangle with three sides of the same length, and an **isosceles triangle** is one with at least two sides of the same length. (Thus by our definition, an equilateral triangle is isosceles.) A quadrilateral whose angles are all right angles is called a **rectangle**. A rectangle whose sides are all of the same length is called a **square**. Be aware that at this point, we do not *know* whether there is a square or not, or worse, whether there is a rectangle or not even if common sense says there is. (If it is the case that the sum of (the degrees) of the four angles of quadrilateral is $361°$, then clearly no rectangle can exist, much less a square.) A quadrilateral with at least one pair of opposite sides that are parallel is called a **trapezoid**. A trapezoid with two pairs of parallel opposite sides is called a **parallelogram**. A quadrilateral with four sides of equal length is called a **rhombus**. Although we will not have time to prove it, it is a fact that a rhombus is a parallelogram.

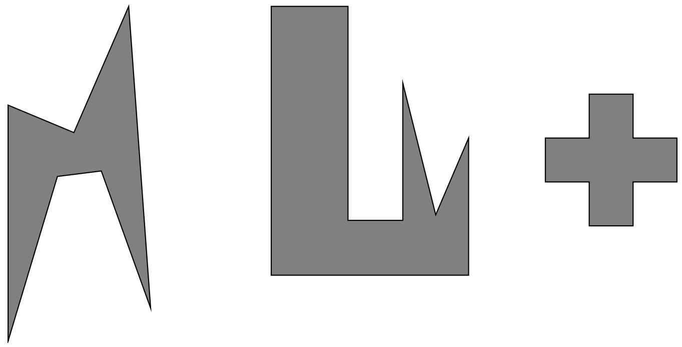
There is a debate in school mathematics about whether one should define an isosceles triangle to have *exactly* two sides of equal length, a rectangle to be a quadrilateral with four right angles but with at least two unequal adjacent sides, or a trapezoid to be a quadrilateral with *exactly* one pair of parallel sides. This debate is unnecessary. Mathematicians want equilateral triangles to be special cases of isosceles triangles, squares to be special cases of rectangles, and parallelograms to be special cases of trapezoids, because this allows every theorem about isosceles triangles to apply to equilateral triangles, every theorem about rectangles to apply to squares, and every theorem about trapezoids to apply to parallelograms. It is completely counterproductive for schools to try to prepare students for advanced work on the one hand, and simultaneously undercut this effort by feeding them erroneous information on the other. There is no reason for school mathematics to make things up when such inventiveness is not called for. Perhaps part of these counter-productive actions are the result of the isolation of school mathematics from the mathematics mainstream. If so, then this is the time to rejoin the mainstream.

[51] The other reason, one may surmise, is students' failure to memorize the new vocabulary.

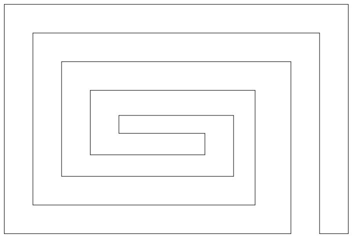
In the above catalog of names for polygons, we all know that equilateral triangles and squares are special; they are examples of *regular polygons*. It turns out that the precise definition of the latter concept is subtle and requires an elaborate discussion. For the needs of middle school mathematics, it would not be profitable to spend time on such subtleties. We will therefore cut the Gordian knot by assuming that, given a polygon $\mathcal{P}$, you know what is meant by **the region enclosed by $\boldsymbol{\mathcal{P}}$**. For example, if we have polygons such as the following:



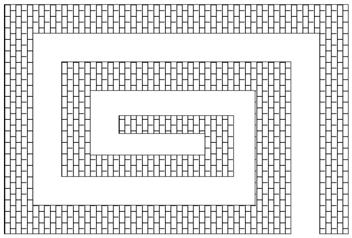
then the regions they enclose are easily recognized as



In case you wonder why the concept of the region enclosed by a polygon is considered to be subtle, there are polygons for which it is not clear what region it encloses. For example, take a point near the center of the following figure, is it clear whether it is or is not in the region enclosed by the polygon?



One can of course carefully make out that the region in question to be the shaded one below:



Nevertheless, one now recognizes that there are far more complicated polygons of this genre, and this explains the subtlety.

One notes also that the concept of the region enclosed by a polygon is entirely analogous to the earlier one of the region enclosed by a circle, except that, by accident, the latter is easy to define.

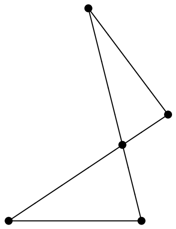
Finally, we are in a position to define regular polygons. We say a polygon is **convex** if the polygonal region enclosed by the polygon is a convex set. A **regular polygon** is a convex polygon which has the property that its sides are of the same length and its angles (at the vertices) have the same degrees. We call attention to the fact that, without the condition of the convexity of the polygon, the cross on the right in the previous picture of three polygons would be a regular 12-gon because of our standing convention that an angle is always automatically taken to be the convex set determined by the two rays. It goes without saying that the convexity of a polygon cannot be defined without the concept of *the region it encloses.*[52]

[52] On the other hand, if one is willing to spend the effort, one *can* define the concept of an *interior angle* of a polygon at some cost. Then a regular polygon could be equally well defined as any polygon whose sides all have equal lengths, and all of whose interior angles have the same degrees.

It is well-known that for a triangle (a 3-gon) to be regular, it suffices to require either the equality of the (lengths of the) sides or the equality of the (degrees of the) angles. For this reason, a regular 3-gon is just an equilateral triangle (which literally means a triangle with equal sides). Moreover, as soon as we can show that the sum of all the angles of a convex quadrilateral is 360°, it would follow that a regular 4-gon must be a square. It can be proved that regular n-gons exist for any whole number $n \geq 3$, but this is certainly not obvious.

Exercises 5.1

1. Explain clearly why the following figure with *five* vertices, as indicated, cannot be a polygon. (*Caution:* Be careful. This is harder than you think.)



2. The intersection of a finite number of convex sets is convex. (The restriction of finiteness is unnecessary, but that proof requires some fluency in dealing with infinite sets and may not be appropriate for middle school.)

3. (a) Suppose we have a finite or an infinite number of convex sets $\mathcal{C}_i$, where i is a whole number, and suppose that each $\mathcal{C}_i$ is contained in the next one, $\mathcal{C}_{i+1}$. Then the union of these $\mathcal{C}_i$'s is also convex. (*Caution*: be very clear in your proof.) (b) Is the union of convex sets convex in general?

4. (a) Prove that half-lines and half-planes are convex. (b) Prove that *closed* half-planes are also convex.

5. Explain why, given any three non-collinear points A, B, C, the three segments AB, BC, CA can never intersect each other except at the endpoints as indicated.

(In other words, take any three noncollinear points, then the union of the segments joining them is always a polygon. It is even a convex polygon; see the next problem.)

6. For a triangle, its **triangular region** can be precisely defined as the intersection of its three angles. (a) Show that the triangular region of a triangle is always convex. (b) Explicitly exhibit the triangular region as the intersection of *three* closed half-planes.

7. (a) Suppose we have three distinct lines, L_1, L_2, and L_3, such that $L_1 \parallel L_2$ and $L_1 \parallel L_3$. Then the line L_2 is parallel to the line L_3, (This problem justifies the terminology that **three lines are parallel**.) (b) Let $L_1 \parallel L_2$ and let a third line ℓ be distinct from L_1. If ℓ intersects L_1, then it must intersect L_2.

8. Imagine the hands of a clock to be idealized rays emanating from the center of the clock. What is the angle between the hands at 8:20?[53]

9. Given a circle C and a point P on C. A line L_P is said to be a **tangent** to C at P if L_P intersects C exactly at P, i.e., $L_P \cap C = \{P\}$. Assume that every point of a circle has a tangent, and furthermore that the circle always lies entirely in one of the closed half-planes of each tangent. Then prove that the circular region of a circle is always convex. (*Caution*: This is not an easy problem; try to do everything according to the definitions.)

10. Let $\triangle A'B'C'$ be inside (the closed region enclosed by) $\triangle ABC$. Assuming the triangle inequality (D4), this problem proves that the perimeter of $\triangle A'B'C' \leq$ the perimeter of $\triangle ABC$. (a) Prove that such is the case if, in addition, two vertices of $\triangle A'B'C'$ lie on one side of $\triangle ABC$. (b) Prove that such is also the case if, in addition, two of the vertices of $\triangle A'B'C'$ lie on two different sides of $\triangle ABC$. (c) Prove the general case. (*Hint:* Let the rays $R_{A'B'}$ and $R_{A'C'}$ meet the sides of $\triangle ABC$ at D and E and consider $\triangle ADE$. Try to apply part (a) or part (b).)

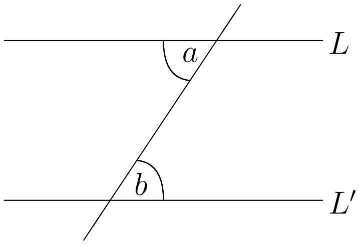
[53] Problem due to Tony Gardiner.

2 Transformations of the plane

In this section, we continue our discussion of the geometry of the plane by defining the abstract concept of a transformation in order to put on a solid foundation what we did in Chapter 4 about "moving one transparency against another." As we shall see, we will be concerned only with a special class of transformations. The *mathematics* of the discussion is conceptually self-contained, but for illustrative purposes — and for those purpose only — we will freely avail ourselves of the discussion in Chapter 4.

Given two segments AB and CD (in the plane), how can we find out if they have the same length without measuring them individually? For example, suppose we have a rectangle $ABCD$. Are the opposite sides of equal length?

Similarly, given two angles, how can we tell whether they are equal without actually measuring the angles individually? For example, if two lines L and L' are parallel and they are intersected by another line, how can we tell if the angles $\angle a$ and $\angle b$ as shown have the same degrees?



These questions, while seemingly silly if we only look at figures on a piece of paper, take on a new meaning if the sides of the rectangle $ABCD$ are several miles apart, or if the lines L and L' are also very far apart. We are therefore confronted with a real-world situation of having to find out whether two general geometric figures (two segments, two angles, or two triangles) in different parts of the plane (possibly very far apart) are "the same" in some sense (e.g., same length, same degrees, etc.).

The traditional way of dealing with this problem in Euclidean geometry is to write down a set of axioms which abstractly guarantee that the two figures in question are "the same" (i.e., congruent). This is how it is usually done in the school classroom, and the drawback of such an approach is that it is an abstraction imposed from on high and students do not have the needed foundations to handle the drastic change in methodology. Consequently, the logical deductions from axioms cease to hold any meaning and proofs become a futile exercise in formalism. As a reaction, the recent trend is to ignore the fundamental deductive nature of mathematics and let students approach geometry exclusively through hands-on activities. At the moment (as of 2010), the teaching of geometry in high schools vacillates between these two extremes, neither of which is designed to encourage the learning of mathematics.[54]

We propose a third alternative by adopting an approach that is more direct, more tangible and makes use of three standard "motions" of Chapter 4 to bring one figure on top of another in order to check whether the two geometric figures are "the same". Even more importantly, we base proofs of theorems directly on these "motions". In this way, the concept of congruence ceases to be abstract and intangible; it can be realized concretely. So the key idea is how to "move" things around in a plane, as we already saw in Chapter 4. However, we used the intuitive language of "rules" and "moving one point to another" in that chapter, and now is the time to adopt a more civilized language that is universally used in mathematics, the language of transformations and mappings.

The rest of this section gives a short introduction to transformations so that we can discuss, in the next section, the basic isometries introduced in Chapter 4 with precision and with ease.

For convenience, we denote the plane by Π. A **transformation** F of Π is a rule that assigns to each point P of Π a unique point $F(P)$ (read: "*F of P*") in Π, so

[54] An overview of the situation can be glimpsed in the book review of H. Wu, Geometry: Our Cultural Heritage, *Notices of the American Mathematical Society*, 51 (2004), 529-537 (available at http://math.berkeley.edu/∼wu/Holme3.pdf). See also H. Wu, Euclid and high school geometry, http://math.berkeley.edu/∼wu/Lisbon2010_1.pdf.

that,

> (i) if P_1 and P_2 are distinct points of Π, then F assigns to them distinct points $F(P_1)$ and $F(P_2)$ of Π, and
>
> (ii) for every point Q of Π there is a point P of Π so that F assigns P to Q, i.e., $Q = F(P)$ for some point P.

If a point Q satisfies $Q = F(P)$ for some point P in Π, we also say ***F* maps *P* to *Q***, or that ***Q* is the image of *P* under *F***. An alternate way of expressing (i) is therefore that F maps distinct points to distinct points; in pictorial language, such an F does not "collapse" distinct points into the same point. The second condition (ii) says that no matter what the point Q may be, it is always equal to the image of some P under F. This fact can be expressed in a different way, which may be more intuitive. Given a set $\mathcal{S}$ in the plane Π, we will denote by $\boldsymbol{F(\mathcal{S})}$ the collection of *all* the points $F(P)$, where P runs through all the points in $\mathcal{S}$. We call $F(\mathcal{S})$ **the image of $\boldsymbol{\mathcal{S}}$ under *F***. We also say ***F* maps $\boldsymbol{\mathcal{S}}$ to $\boldsymbol{F(\mathcal{S})}$.** In this terminology, condition (ii) may be rephrased as saying that the image of the plane under F, $F(\Pi)$, is the whole plane Π.

In the mathematics literature, transformations are not required to satisfy either of the two conditions (i) and (ii). The common terminology to express (i) is that the transformation F is **one-to-one**, and that to express (ii) is that F is **onto**. A transformation that satisfies both (i) and (ii) is usually called a **one-one correspondence**. Thus the transformations in these notes are, *by definition*, one-one correspondences. For our purposes, it is sufficient to consider only one-one correspondences.

Informally, the dilations of §4 of Chapter 4 are examples of transformations, as are the three basic isometries, translations, reflections, and rotations. However, the three basic isometries differ from dilations in that they *preserve distance*. Formally, an **isometry** of the plane is a transformation that preserves the lengths of segments. Or, an isometry F is a transformation so that the length of any segment is equal to the length of its image segment. In other words, if F is an isometry and if we denote the image points of P and Q under F by P' and Q', then the length of the segment

$P'Q'$ is always equal to the length of the segment PQ for any points P and Q. In terms of the concept of distance on Π introduced in §1, we see that for an isometry F,

$$\operatorname{dist}(F(P), F(Q)) \; = \; \operatorname{dist}(P, Q) \quad \text{for all } P, Q \in \Pi.$$

It is in this sense that **isometries are transformations that preserve distance.**

There is one isometry that is distinguished: the **identity transformation $\boldsymbol{I}$** so that $I(P) = P$ for all P in the plane. Unless stated to the contrary, I will always stand for the identity transformation.

Next, we introduce a few more concepts. In Chapter 4, we saw many examples of combining the actions of two basic isometries one after the other. For example, if ρ is the rotation of 15 degrees around a point O and if ρ^* is the rotation of 22 degrees around the same point O, then we know that moving a point P first by ρ, and then by ρ^* has the net effect of rotating P by a total degree of $15 + 22$ around O. You recognize that if we combine isometries this way often, such verbal descriptions will get too cumbersome very fast. We have to formalize both the concept and the notation in order to streamline the discussion.

Formally, If F and G are transformations, we say **the transformations $\boldsymbol{F}$ and $\boldsymbol{G}$ are equal**, in symbols $\boldsymbol{F = G}$, if $F(Q) = G(Q)$ for *every* point $Q \in \Pi$.[55] The **composite transformation $\boldsymbol{F \circ G}$** (sometimes also called the **composition of $\boldsymbol{F}$ and $\boldsymbol{G}$**) is by definition the transformation which assigns a point P in the plane to the point $F(G(P))$, i.e., $(F \circ G)(P) \overset{\text{def}}{=} F(G(P))$ for all $P \in \Pi$. For example, no matter what F is, we will always have $F \circ I = I \circ F = F$. Of course we should double-check that $F \circ G$ so defined *is* a transformation, i.e., satisfies conditions (i) and (ii). This is straightforward and involves nothing more than going through the definition methodically, and is best done by you as an exercise to get acquainted with the new definitions.

To go back to the previous rotations of ρ and ρ^*, in terms of the notation of composition, we can now write the combined action as $(\rho^* \circ \rho)(P)$. Notice that in

[55] Observe that we have now introduced a new meaning to the equal sign, the equality of two transformations. This is different from everything we have done up to this point, because in the past we have only used the equal sign between two numbers or two sets. Since this definition is completely unambiguous, however, there is no need to guess what "equal" means in this case.

order to express "ρ first", we have to place ρ right next to P and therefore ρ ends up being to the right rather than to the left of ρ^*.

Note that if F and G are isometries, so is $F \circ G$. Indeed, G being an isometry means, for any two points P and Q, $\text{dist}(G(P), G(Q)) = \text{dist}(P, Q)$. But F is also an isometry, so $\text{dist}(F(G(P)), F(G(Q))) = \text{dist}(G(P), G(Q))$. Together, we have that for any two point P and Q,

$$\text{dist}((F \circ G)(P), (F \circ G)(Q)) \;=\; \text{dist}(F(G(P)), F(G(Q))) \;=\; \text{dist}(P, Q)$$

This then shows that $F \circ G$ is an isometry.

Now we come to the main point: every transformation has another transformation that "reverses" its action, and vice versa. For example, consider a rotation ρ with center O and degree e. The rotation ρ' with the same center O but with degree $-e$ has the property that if $\rho(P) = Q$, then $\rho'(Q) = P$, and also if $\rho'(P) = Q$, then $\rho(Q) = P$. One can summarize the situation better with a pair of equations:

$$\rho(\rho'(P)) \;=\; P \quad \text{and} \quad \rho'(\rho(P)) \;=\; P \quad \text{for any point } P.$$

Recalling that $I(P) = P$ for every point P, these equations can be rewritten as

$$\rho \circ \rho' \;=\; I \quad \text{and} \quad \rho' \circ \rho = I$$

Activity (a) For a translation T from A to B, what is the transformation T^* so that $T \circ T^* = I$ and $T^* \circ T = I$? (b) Same question for a reflection R.

Given a transformation F in general, suppose there is a transformation G so that both $F \circ G = I$ and $G \circ F = I$, where I is the identity transformation of the plane. Then we say G is the **inverse** transformation of F (and of course, also that F is the **inverse transformation of G**). Often, we simply say F **is the inverse of** G. We have the following theorem.

Theorem *Every transformation has an inverse transformation.*

We wish to be explicit about the fact that this theorem is true only because a transformation is, by our definition, both one-to-one and onto.

Proof Let F be a transformation. We now define a new transformation G, as follows. Given a point P, G assigns the point Q to P so that, Q is the point guaranteed by condition (ii) above so that $F(Q) = P$. Now we have to make sure G is well-defined (see the second sub-section in §4 of Chapter 1), i.e., insofar as the definition of a transformation requires that it assigns to each P a unique $Q = F(P)$, we must ask whether there is only one such Q. The answer is affirmative, because if there is another point Q_0, $Q_0 \neq Q$, so that $F(Q_0) = P$, then condition (i) would be violated (we have just "collapsed" two distinct points Q and Q_0 into the same point P). Therefore, there is no such Q_0. So G is a well-defined transformation.

Now we must prove that the transformation G satisfies $G \circ F = F \circ G = I$. We first check $G \circ F = I$. Take a point P, and we have to prove $G(F(P)) = P$. Let $Q = F(P)$, then by the definition of G, we have $G(Q) = P$. But this is exactly the statement that

$$G(F(P)) \;=\; G(Q) \;=\; P$$

So $G \circ F = I$. Next we check why $F \circ G = I$. For each point P, we have to prove $F(G(P)) = P$. Let $G(P) = Q$. By the definition of G, G assigns Q to P because $F(Q) = P$. It follows that $F(G(P)) = F(Q) = P$. The proof of the theorem is complete.

It remains to point out that the standard notation for the inverse transformation of a transformation F is $\boldsymbol{F^{-1}}$.

Exercises 5.2

1. If a transformation maps lines to lines, then it maps a pair of parallel lines to a pair of parallel lines.

2. Prove that if F and G are transformations of the plane, then so is $F \circ G$, i.e., $F \circ G$ satisfies conditions (i) and (ii).

3. We spoke freely in the text about "the" inverse transformation of a given transformation. This problem legitimizes this language by asking you to prove that the

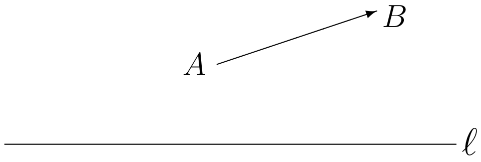
inverse of a transformation F, if it exists, must be unique. In other words, suppose there are two transformations G and G' so that

$$\begin{aligned} F \circ G &= I \quad \text{and} \quad G \circ F = I \\ F \circ G' &= I \quad \text{and} \quad G' \circ F = I \end{aligned}$$

Then $G = G'$ (recall: this means $G(P) = G'(P)$ for every point P in the plane.)

4. Do this problem using the information of Chapter 4. (a) What is the inverse of a rotation of 180 degrees (with any center)? (b) Express a rotation of 180 degrees as the composition of two reflections.

5. (a) Let T be the translation from A to B as shown, and let R be the reflection across the line ℓ as shown. Using the information of Chapter 4, what is the inverse transformation of $R \circ T$, and what is the inverse transformation of $T \circ R$?



(b) Let F, G be transformations, and let F^{-1} and G^{-1} be their respective inverse transformations. What are the inverse transformations of $F \circ G$ and $G \circ F$ in terms of F^{-1} and G^{-1}.

3 The basic isometries: Part II

Rotations

Reflections

Translations

We will formally introduce in this section the three **basic isometries** of the plane, namely, rotations, reflections, and translations, and make a few tentative steps toward proving theorems in geometry. We do geometric proofs only selectively at this point. Our goal, after all, is to do enough geometry so that we can discuss the geometry of linear equations with ease, and a full-blown course on geometry at this point will be a distraction. On the other hand, one cannot present mathematical facts without at least some reasoning to hold them together. So we will present proofs that are truly basic and instructive.

Please keep in the back of your mind our intuitive discussions of the basic isometries in Chapter 4.They will be very helpful in providing support for the understanding of the formal mathematics. On the other hand, the following *formal* discussion does not *logically* assume anything from Chapter 4, in the sense that none of the logical arguments in a proof will make use of any conclusions arrived at by hands-on experiments alone.

We continue to denote the plane by Π.

Rotations

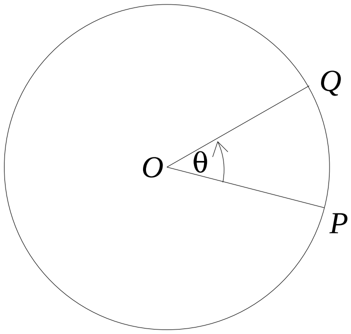
Let O be a point in the plane and let a number θ be given so that $-180 \leq \theta \leq 180$. Notice that we allow θ to be both 180 and -180. Then the **rotation of $\boldsymbol{\theta}$ degrees around $\boldsymbol{O}$** (or sometimes we say **with center $\boldsymbol{O}$**) is the transformation $\boldsymbol{\rho_\theta}$ defined as follows: $\rho_\theta(O) = O$, and if $P \in \Pi$ and $P \neq O$, let C be the circle of radius $|OP|$ centered at O; then

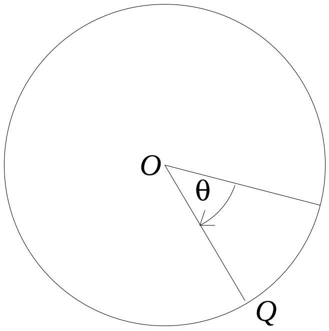
> $\rho_\theta(P)$ is the point Q on C so that if $\theta \geq 0$, Q is in the counterclockwise direction of P along C and $|\angle QOP| = \theta^\circ$, and if $\theta < 0$, Q is in the clockwise direction of P along C and $|\angle POQ| = |\theta|^\circ$.

The case of $\theta > 0$ is illustrated by the following figure:
while the case of $\theta < 0$ looks like this:

Before proceeding further, we must prove that ρ_θ *so defined is indeed a transformation, i.e., it satisfies conditions* (i) *and* (ii) *of* §*2.*

First, condition (i). Let P and M be distinct points in the plane and let $\rho_\theta(P) = Q$ and $\rho_\theta(M) = N$. Then we have to prove $Q \neq N$. If $|OP| \neq |OM|$, then there is nothing to prove because, by definition, $|OP| = |OQ|$ and $|OM| = |ON|$, so that $|OQ| \neq |ON|$ and therefore $Q \neq N$. Now if $|OP| = |OM|$, then both P and N are





distinct points on the same circle C around O of radius $|OP|$. Because ρ_θ rotates both P and M along C by θ degrees (clockwise or counterclockwise, depending on whether θ is negative or positive), Q and N remain distinct. So again $Q \neq N$, and condition (i) is satisfied.

Before verifying that ρ_θ satisfies condition (ii), we make some useful general observations about rotations. Note that the rotation of 0 degrees around a point is just the identity transformation I of the plane. If θ and θ' are numbers so that $-180 \leq \theta, \theta' \leq 180$ and also $-180 \leq \theta + \theta' \leq 180$, then relative to the *same* center of rotation, the composition of the rotations ρ_θ and $\rho_{\theta'}$ can be seen to satisfy

$$\rho_\theta \circ \rho_{\theta'} = \rho_{\theta+\theta'}$$

In particular, for any such θ,

$$\rho_\theta \circ \rho_{-\theta} = I \quad \text{and} \quad \rho_{-\theta} \circ \rho_\theta = I$$

Therefore $\rho_{-\theta}$ is the inverse transformation of ρ_θ for $-180 \leq \theta \leq 180$.

Now we can dispatch condition (ii) of §2. Given a point Q in the plane, we have to find a point P so that $\rho_\theta(P) = Q$. We simply let $P = \rho_{-\theta}(Q)$. Then

$\rho_\theta(P) = \rho_\theta(\rho_{-\theta}(Q) = Q$, and we have verified that ρ_θ is indeed a transformation.

Notice that, by choice, we have allowed the two transformations ρ_{-180} and ρ_{180} to be the same transformation. This will not cause confusion.

We make the following assumptions about rotations:

($\boldsymbol{\rho}$1) Given any point $O \in \Pi$ and any number θ satisfying $-180 \leq \theta \leq 180$, there is a rotation of θ degrees around O.

($\boldsymbol{\rho}$2) Any rotation maps a line to a line, a ray to a ray, and therefore a segment to a segment.

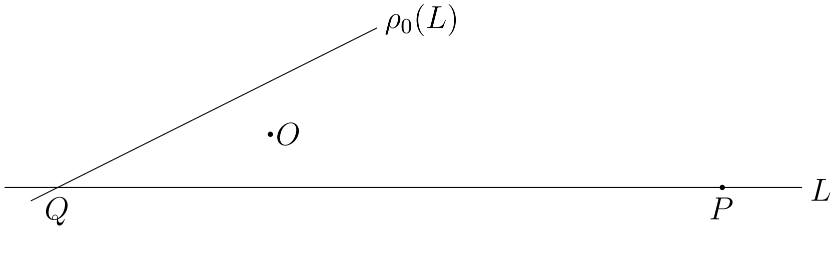
($\boldsymbol{\rho}$3) Any rotation preserves length of segments (and is therefore an isometry) as well as degrees of angles.

These are all believable in view of the hands-on activities we did in Chapter 4, §3.

Having rotations at our disposal, our next goal is to define the other basic isometries. To this end, we need to know some basic facts about perpendicular lines and, therefore, have no choice but to begin proving theorems. We will not indulge in theorem-proving per se, but will only prove enough to define reflections and translations. Later on, we will just prove enough to understand similarity.

We should point out that, henceforth, all the theorems on plane geometry will be numbered consecutively by G1, G2, G3, etc., so that you will know how to get started if you decide to give a coherent account of plane geometry.

Theorem G1 *Let O be a point not contained in a line L, and let ρ_0 be the rotation of* $180°$ *around O. Then ρ_0 maps L to a line parallel to itself, i.e., $\rho_0(L) \parallel L$.*



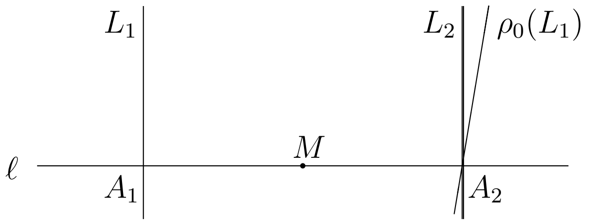
Before we get started on the proof, it will be helpful if you remember the discussion of questions (i) and (ii) at the end of §3 in Chapter 4. There we actually *proved* that if P is a point on L, then $\rho_0(P)$ cannot lie on L. The whole proof of Theorem G1 hinges on this fact. Another thing to note in general is that, if you are ever asked to prove that a line ℓ is parallel to another line L, then you know that, in most cases, a direct proof (one that manages to deduce directly that ℓ does *not* intersect L) is well-nigh impossible. You may wish to try right away a proof by contradiction: assume that ℓ intersects L and then deduce an absurd conclusion. The absurdity then implies that you shouldn't have made that assumption in the first place. So ℓ does not intersect L.

We have used this kind of proof before, e.g., in §3 of Chapter 4. Students tend to be puzzled by proof by contradiction. Nevertheless, it won't hurt you to try it a few times; even if they don't understand it in middle school, the exposure may help them understand it later. Sometime the gestation period for understanding is long, but it is a worthwhile attempt to try to win the war even if you know you will lose a battle or two.

Proof Suppose $\rho_0(L)$ is not parallel to L. We will show that this assumption leads to a conclusion that contradicts the hypothesis that P does not lie on L. So $\rho_0(L)$ is not parallel to L and they intersect at a point Q. Since $Q \in \rho_0(L)$, by definition of $\rho_0(L)$, there is a point $P \in L$ so that $\rho_0(P) = Q$. Since ρ_0 is a rotation of 180° around O, the three points P, O, and $\rho_0(P)$ are collinear, i.e., P, O, and Q are collinear. As usual, call this line L_{PQ}. Now, not only is P on L, but Q is also on L because $Q = L \cap \rho_0(L)$. Thus L and L_{PQ} have two points P and Q in common and therefore they coincide: $L = L_{PQ}$ (see (L1) of §1). But O also lies on L_{PQ}, so O lies on L, and this directly contradicts the hypothesis that O is not contained in L. Therefore $\rho_0(L)$ has to be parallel to L.

Theorem G2 *Two lines perpendicular to the same line are either identical or parallel to each other.*

Proof Let L_1 and L_2 be two lines perpendicular to a line ℓ at A_1 and A_2, respectively. If $A_1 = A_2$, then as we noted in §1 about the uniqueness of the line passing through a given point of a line and perpendicular to that line, L_1 and L_2 are identical. So we may assume that $A_1 \neq A_2$. We need to prove that $L_1 \parallel L_2$. We have just seen how to produce a line parallel to a given line using a rotation of 180 degrees, so we should at least exploit this fact. Let ρ_0 be the rotation of 180 degrees around the midpoint M of A_1A_2. If we can show that the image of L_1 under ρ_0 is L_2, then we know $L_2 \parallel L_1$ by virtue of Theorem G1.



To this end, note that $\rho_0(L_1)$ contains A_2 because $\rho_0(A_1) = A_2$ (ρ_0 is an isometry). We are given that $L_1 \perp \ell$. Since ρ_0 is a rotation of 180 degrees and the center of rotation, which is M, lies on ℓ, we see that $\rho_0(\ell) = \ell$. By property (ρ3), which implies that rotations map perpendicular lines to perpendicular lines, we have $\rho_0(L_1) \perp \ell$. Therefore each of $\rho_0(L_1)$ and L_2 is a line that passes through A_2 and perpendicular to ℓ. Because of the uniqueness of the line perpendicular to a given line at a given point, we see that, indeed, $\rho_0(L_1) = L_2$. By Theorem G1, $L_2 \parallel L_1$, and Theorem G2 is proved.

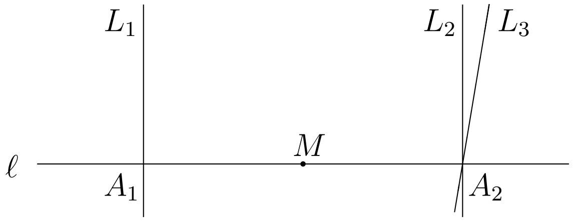
Recall that we have introduced the concept of a rectangle as a quadrilateral whose adjacent sides are all perpendicular to each other. As a result of Theorem G2, we now have:

Corollary *A rectangle is a parallelogram.*

Given two lines L_1 and L_2. A **transversal** of L_1 and L_2 is a line ℓ that intersects both in distinct points. The following fact rounds off the picture of Theorem G2, and is also hinted at in the discussion (in §4 of Chapter 4) of the lines on your notebook papers.

Theorem G3 *A transversal of two parallel lines that is perpendicular to one of them is also perpendicular to the other.*

Proof Let $L_1 \parallel L_2$ and let the transversal ℓ meet L_1 and L_2 at A_1 and A_2, respectively. Assuming $L_1 \perp \ell$, we will prove that $L_2 \perp \ell$.

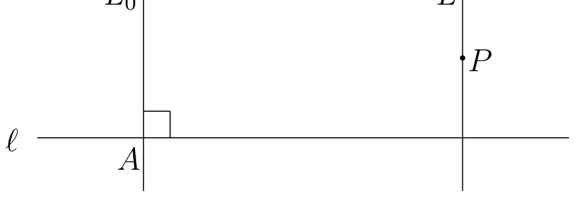


Let L_3 be a line perpendicular to ℓ at A_2. (Recall that L_3 is the angle bisector of the straight angle with sides lying on ℓ and with vertex at A_2.) By Theorem G2, $L_3 \parallel L_1$. But by hypothesis, $L_2 \parallel L_1$. The Parallel Postulate therefore implies that $L_3 = L_2$. Since $L_3 \perp \ell$, we now have $L_2 \perp \ell$. This proves Theorem G3.

For our immediate need in connection with the definition of a reflection, the following consequence of Theorem G3 should be singled out:

Corollary *Through a point P not lying on a line ℓ passes one and only one line L perpendicular to ℓ.*

Proof Take any point $A \in \ell$ and let L_0 be the line passing through A and perpendicular to ℓ. By the Parallel Postulate, there is a line L passing through P and parallel to L_0.



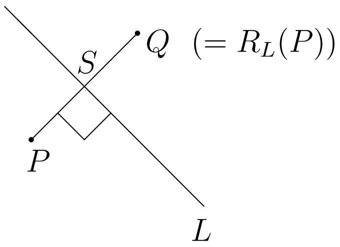
By Theorem G3, we have $L \perp \ell$. To prove the uniqueness of L, suppose another line L' passes through P and is also perpendicular to ℓ. By Theorem G2, since these lines are not parallel (because they have P in common), they have to be identical. Thus $L = L'$.

Reflections

We can now define reflection. Given a line L, the **reflection across $\boldsymbol{L}$** (or **with respect to $\boldsymbol{L}$**) is by definition the transformation $\boldsymbol{R_L}$ of Π, so that:

(1) If $P \in L$, then $R_L(P) = P$.

(2) If $P \notin L$, then $R_L(P)$ is the point Q so that L is the perpendicular bisector of the segment PQ.



We hasten to show that the definition is well-defined, in the sense that the assignment of $R_L(P)$ to each point P is a transformation. First, is the assignment $P \mapsto R_L(P)$ unique? For $P \in L$, there is no doubt. Suppose now P is not on L. By the Corollary to Theorem G3, there is a unique line passing through P and perpendicular to L; let us say this line intersects L at S. On the ray R_{PS}, we take a point Q so that Q and P lie in opposite half-planes of L and so that $|PS| = |SQ|$. Then according to (2) above, this Q is $R_L(P)$. Suppose there is another Q' that satisfies (2), i.e., L is the perpendicular bisector of PQ'. Then $PQ \perp \ell$ and $PQ' \perp \ell$; therefore the lines L_{PQ} and $L_{PQ'}$ are the same because of the uniqueness part of the Corollary to Theorem G3. It follows that Q' lies on L_{PQ}, and $Q = Q'$ because Q and Q' are in the same half-plane of L, and $|QS| = |Q'S|$. Thus the definition of a reflection *is* well-defined.

> *This may be the right place to pause and reflect on some of the materials in §1. The preceding discussion draws on the fact that a line divides the plane into two half-planes and that the concept of a ray is clearly understood. If you had any doubts about why we bothered to list* ($L4$) *and* ($L5$) *as part of our starting point, you may lay these doubts to rest now.*

It remains to verify that R_L satisfies the conditions (i) and (ii) in the definition of a transformation. Having done this for rotations, we'd leave this verification as an exercise.

Reflections enjoy a remarkable property. Fix a line L, and let R be the reflection with respect to L. Then it is straightforward to check that $R \circ R = I$, where I as usual denotes the identity transformation of the plane. But this means R *is its own inverse.*

As in the case of rotations, we make the following entirely plausible assumptions about reflections:

($\boldsymbol{R}$1) Given any line L, there is a reflection with respect to L.

($\boldsymbol{R}$2) Any reflection maps a line to a line, a ray to a ray, and therefore a segment to a segment.

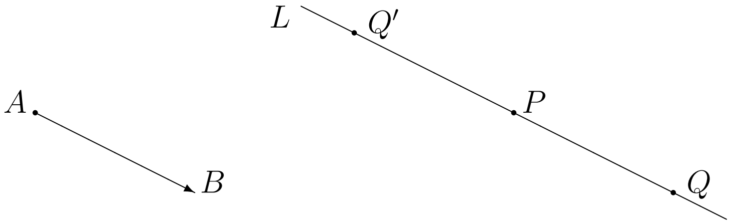
($\boldsymbol{R}$3) Any reflection preserves lengths of segments (and is therefore an isometry) as well as degrees of angles.

We give a simple application of reflections by proving: *every point on the perpendicular bisector of a segment is equi-distant from the endpoints of the segment.*[56] Indeed, let ℓ be the perpendicular bisector of BC and let $A \in \ell$. We have to prove $|AB| = |AC|$. Let R be the reflection with respect to ℓ. By the definition of reflection, we see that $R(B) = C$ and $R(A) = A$. Therefore ($R2$) implies that $R(AB) = AC$. By ($R3$), we have $|AB| = |AC|$.

[56] Compare item 10 in §1 of Chapter 4.

Translations

The last basic isometry to be introduced is translation. Intuitively, the translation T, along the direction from point A to point B and of distance $|AB|$, does the following to an arbitrary point P in the plane: draw the line L passing through P and parallel to L_{AB}, then on the line L, we mark off the point Q so that $|PQ| = |AB|$ and so that the direction from P to Q is the same as the direction from A to B. By definition, $Q = T(P)$.



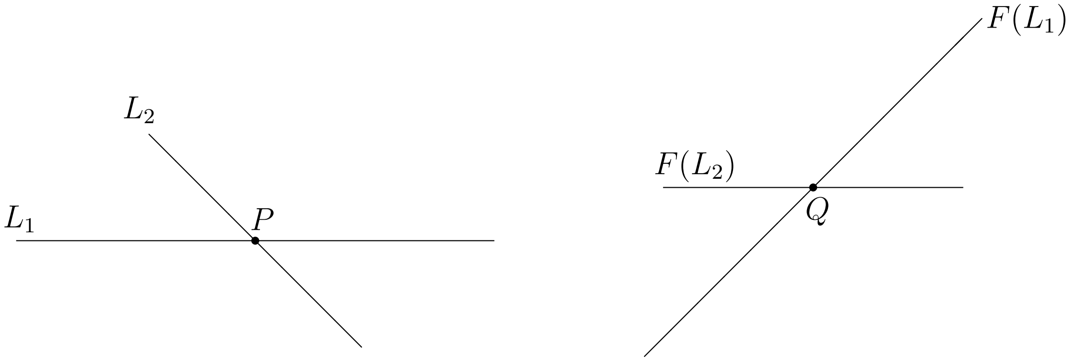
Intuitively, all this is good and well. But in terms of precision, the difficulty with this description lies in the fact that on L, there is a point Q' which is also of distance $|AB|$ from P but the direction from P to Q' is "opposite" to that from P to Q (see picture above). The problem becomes one of how to say, precisely, that it is Q and not Q' that should be defined as $T(P)$. The following discussion is designed to circumvent this difficulty. The trick is to observe that $ABQP$ is a parallelogram but $ABQ'P$ is not. So we begin with a discussion of a key fact about parallelograms that will eventually make the definition of a translation more meaningful.

To this end, it would be helpful to adopt a common abuse of language: we say two segments are **equal** if their lengths are equal, and say two angles are **equal** if their degrees are equal. The following will be helpful in our discussion of translations.

Theorem G4 *Opposite sides of a parallelogram are equal.*

Theorem G4 together with the Corollary to Theorem G2 imply that the *opposite sides of a rectangle are equal.* This reconciles the usual definition in school mathematics of a rectangle (a quadrilateral with four right angles and equal opposite sides) with our definition of a rectangle (a quadrilateral with four right angles). The proof of Theorem G4 requires the following lemma.

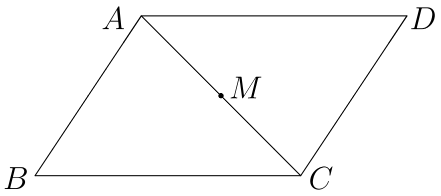
Lemma *If F is a transformation of the plane that maps lines to lines, then for any two distinct lines L_1 and L_2, if $L_1 \cap L_2 = \{P\}$ and $F(L_1) \cap F(L_2) = \{Q\}$, then $F(P) = Q$.*



Proof of Lemma Since $P \in L_1$, we have $F(P) \in F(L_1)$ by the definition of the image $F(L_1)$ of L_1 under F. Similarly, $F(P) \in F(L_2)$. Therefore $F(P)$ is a point in the intersection $F(L_1) \cap F(L_2)$. But by hypothesis, the latter intersection is exactly the point Q. So $F(P) = Q$.

Proof of Theorem G4 This proof, like the proofs of Theorems G1 and G2 and like many others to come, is nothing but an exploitation of the simple fact that a 180 degree rotation is an isometry that also preserves the degrees of angles. In this case, we have a parallelogram $ABCD$. You will see that the 180 degree rotation around the midpoint of a diagonal reveals a surprising amount of geometric information about the parallelogram.

Given parallelogram $ABCD$, we must show $|AD| = |BC|$ and $|AB| = |CD|$. It suffices to prove the former. Let M be the midpoint of the diagonal AC:



Let ρ_0 be the rotation of 180 degrees around M. Then $\rho_0(C) = A$ so that $\rho_0(L_{BC})$ is a line passing through A and (by Theorem G1) parallel to L_{BC}. Since the line L_{AD} has exactly the same two properties by assumption, the Parallel Postulate implies that $\rho_0(L_{BC}) = L_{AD}$. Similarly, $\rho_0(L_{AB}) = L_{CD}$. Thus,

$$\rho_0(L_{BC}) \cap \rho_0(L_{AB}) = L_{AD} \cap L_{CD} = \{D\}$$

On the other hand, $L_{BC} \cap L_{AB} = \{B\}$. By the Lemma, we have

$$\rho_0(B) = D$$

Recall we also have $\rho_0(C) = A$. Therefore $\rho_0(BC) = AD$. Since ρ_0 is an isometry (by (ρ3)), we have $|BC| = |AD|$, as desired.

Remark It may be of interest to add a comment to the preceding assertion that, because $\rho_0(B) = D$ and $\rho_0(C) = A$, we have $\rho_0(BC) = AD$. This is so intuitively obvious that, at least in the school classroom, it would probably be wise to leave it as is. As a teacher, however, you should be aware that it is possible to give a precise proof, which runs as follows. By assumption (ρ2), ρ_0 maps the segment BC to the segment $\rho_0(BC)$ which joins $\rho_0(B)$ and $\rho_0(C)$. Thus $\rho_0(BC)$ is a segment which joins D ($= \rho_0(B)$) to A ($= \rho_0(C)$. But AD is a segment joining D to A, and there is only one segment joining D to A, by (L1). Therefore $\rho_0(BC) = AD$.

Corollary to Theorem G4 *The angles of a parallelogram at opposite vertices are equal.*

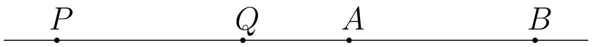
The proof is already implicit in the proof of Theorem G4, and will therefore be left as an exercise.

We are now ready to define translation. We first extend the concept of a *vector* first introduced in §2 of Chapter 2 from the number line to the plane. Given two points A and B in Π, the **vector** $\overline{\boldsymbol{AB}}$ is the segment AB together with a **starting point** A, which is the first letter in $\overline{AB}$, and an **endpoint** B, which is the second letter in $\overline{AB}$.[57] In other words, a vector is just a segment together with a **direction**

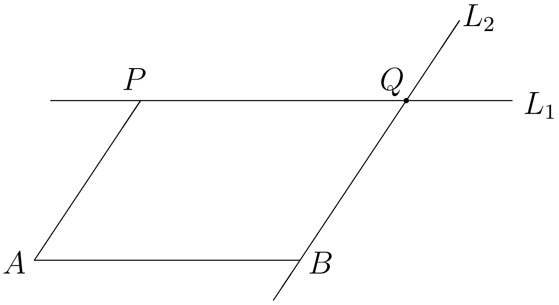
[57] This is the same concept as the one used in calculus on $\mathbf{R}^3$, though the notation may be slightly different.

from a designated endpoint to the other endpoint. For example, while the segments AB and BA are the same, the *vectors* $\overline{AB}$ and $\overline{BA}$ are different because they have different starting points and endpoints. With this understood, given a vector $\overline{AB}$, we define the **translation along $\boldsymbol{\overline{AB}}$** to be the transformation $\boldsymbol{T_{AB}}$ of Π so that:

> (1) If $P \in L_{AB}$, then $T_{AB}(P)$ is the point $Q \in L_{AB}$ so that $\overline{PQ}$ has the same length and same direction as $\overline{AB}$. More precisely, if we regard L_{AB} as the number line *so that the starting point A of $\overline{AB}$ is 0* and so that *the endpoint B is a positive number* (to the right of 0), then Q is the point on L_{AB} also to the right of P and $|PQ| = |AB|$.



> (2) If $P \notin L_{AB}$, then $T_{AB}(P)$ is the point Q obtained as follows. Let L_1 be the line passing through P and parallel to L_{AB}. Let L_2 be the line passing through the *endpoint* B of $\overline{AB}$ and parallel to the line L_{AP}, which joins the *starting point* A of $\overline{AB}$ and P. The point Q is the intersection of L_1 and L_2.



We will sometimes refer to T_{AB} as the **translation from $\boldsymbol{A}$ to $\boldsymbol{B}$.**

A few supplementary comments would make the definition more intuitive. First, why must L_1 and L_2 in (2) intersect? Suppose not, then $L_2 \parallel L_1$. Now B does not lie on L_1 (because $L_1 \parallel L_{AB}$). Thus through B pass two lines parallel to L_1, namely L_2 and L_{AB}. By the Parallel Postulate, $L_2 = L_{AB}$. In particular, $A \in L_2$ and therefore

L_2 intersects L_{AP} at A. This contradicts the fact that $L_2 \parallel L_{AP}$. Therefore L_1 must intersect L_2. Next,

> *if $T_{AB}(P) = Q$, then the distance from P to Q is always equal to the length $|AB|$.*

This is so if $P \in L_{AB}$ (by part (1) of the definition), and if P is not on L_{AB}, by construction, $ABQP$ is a parallelogram and so by Theorem G4, again $|PQ| = |AB|$. Furthermore, $L_{PQ} \parallel L_{AB}$ or $L_{AB} = L_{PQ}$. In other words, a translation "*moves every point in the plane the same distance and in the same direction.*"

Keeping the same notation, we note that if we consider the vector $\overline{BA}$, then for a point $P \notin L_{AB}$, the translation T_{BA} maps the point Q to exactly P because, according to (2), we obtain $T_{BA}(Q)$ as follows: it is the point of intersection of the line passing through Q and parallel to L_{BA} (that would be L_1 again), and the line passing through the *endpoint* A of $\overline{BA}$ and parallel to L_{BQ} (that would be L_{AP}). This point of intersection is of course just P. Therefore for any point P not lying on L_{AB}, we have

$$T_{BA}(T_{AB}(P)) = P$$

By retracing the steps in (1), it is simple to see that the equality $T_{BA}(T_{AB}(P)) = P$ persists even when $P \in L_{AB}$. Therefore we have

$$T_{BA} \circ T_{AB} = I$$

By switching the letters A and B, we obtain

$$T_{AB} \circ T_{BA} = I$$

This means that *for any vector $\overline{AB}$, the inverse transformation of the translation T_{AB} is T_{BA}.*

We proceed to make the same assumptions about translations as the rotations and reflections:

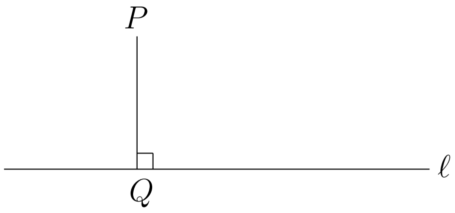
(*T*1) Given any vector $\overline{AB}$, there is a translation along $\overline{AB}$.

(*T*2) Any translation maps a line to a line, a ray to a ray, and therefore a segment to a segment.

(*T*3) Any translation preserves lengths of segments (and is therefore an isometry) as well as degrees of angles.

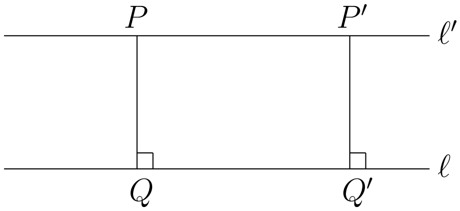
Translations have a noteworthy property: *the translation T_{AB} maps a line L which is neither parallel to L_{AB} nor equal to L_{AB} to a line parallel to L itself.* Suppose not. Then L intersects $T_{AB}(L)$ at a point Q. Since $Q \in T_{AB}(L)$, there is a point $P \in L$ so that $T_{AB}(P) = Q$. But Q is also in L, so P and Q are both in L, and therefore $L_{PQ} = L$. If P lies in L_{AB}, then by part (1) of the definition of a translation, Q is also in L_{AB} and therefore $L = L_{PQ} = L_{AB}$, and this contradicts the hypothesis that $L \neq L_{AB}$. Thus P does not lie in L_{AB}, and by part (2) of the definition of a translation, we have $L_{PQ} \parallel L_{AB}$. Therefore $L \parallel L_{AB}$ and this again contradicts the hypothesis that L is not parallel to L_{AB}. So L is parallel to $T_{AB}(L)$ after all.

We take this opportunity to make a useful observation. Given two parallel lines, we can now define the distance between them. First, let P be a point not lying on a line ℓ. The **distance of *P* from the line *ℓ*** is by definition the length $|PQ|$, where Q is the point intersection of the line ℓ and the line passing through P and perpendicular to ℓ.



Now suppose we have parallel lines ℓ and ℓ', and $P \in \ell'$. If P' is another point on ℓ', then we claim that *the distance of P from ℓ is the same as the distance of P' from ℓ.*[58] Indeed, let the line passing through P' and perpendicular to ℓ intersect ℓ at Q'. By Theorem G2, $L_{PQ} \parallel L_{P'Q'}$. Therefore $PQQ'P'$ is a parallelogram. Consequently, $|PQ| = |P'Q'|$, by Theorem G4. This proves the claim.

[58] This explains why the sleepers (cross ties) across rail tracks can afford to be all of the same length.



The common distance from points on one of two parallel lines to the other is called **the distance between the parallel lines**.

Exercises 5.3

1. (a) Check that a reflection as defined in this section satisfies conditions (i) and (ii) in the definition of a transformation. (b) Do the same for translation.

2. In our definition of translation, we drew a picture to show, if $T_{AB}(P) = Q$, where the point Q is. Using exactly the same notation and same picture, show where Q' is if $T_{BA}(P) = Q'$.

3, Let L be a line in the plane, which may be taken to be the usual number line. Denote 0 on L by A, and the number 1 on L by B. Let ρ_1 be the counterclockwise rotation of 45° around A, and let ρ_2 be the clockwise rotation of 90° around B. Describe as precisely as you can the line $\rho_1\rho_2(L)$ and the line $\rho_2\rho_1(L)$. In particular, does $\rho_1\rho_2(L)$ equal $\rho_2\rho_1(L)$?

4. Prove the Corollary to Theorem G4.

5. If $ABCD$ is a parallelogram, prove that $\angle ADB$ and $\angle CBD$ are equal. (*Caution:* We have not yet proved anything about "alternate interior angle", so you cannot use that fact for this proof. Look instead at the proof of Theorem G4.)

6. Let R be the reflection across the line ℓ. Prove that the image under R of a line parallel to ℓ is another line parallel to ℓ. Prove also that if a line L is not parallel to ℓ, then $R(L)$ always intersects L, in fact, L and $R(L)$ intersect on the line of reflection ℓ.

7. Explain why, given any three non-collinear points in the plane, there is always a circle passing through them.[59] (*Caution:* Rest assured that this does not tax your knowledge of high school geometry, but *only* your understanding of things we have done in this section. Be sure you justify every step. For example, why is it necessary to assume that the three points are not collinear?)

8. Assuming that the lines on your notebook are equi-distant parallel lines, prove that these lines intercept segments of the same length on any transversal. (Compare problem 7(b) in Exercises 5.1.)

9. Given a line L, prove that all the points of a fixed distance k from L form two lines each parallel to L.

4 Congruence

The definition

We begin with a key definition.

Definition *A transformation of the plane* Π *is called a* **congruence** *if it is a composition of a finite number of basic isometries.*

Note that since each basic isometry is a transformation, and compositions of transformations are transformations, a congruence is in particular a transformation. Congruence is one of the main concepts in the school geometry curriculum. Here are its most basic properties:

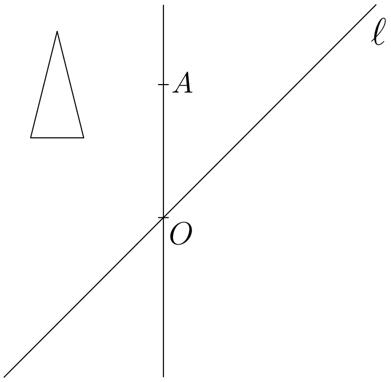
[59] Compare item 13 in §1 of Chapter 4.

Theorem G5 (a) *Every congruence is an isometry that preserves the degrees of angles and maps lines to lines.* (b) *The inverse of a congruence is a congruence.* (c) *Congruences are* **closed under composition** *in the following sense: if F and G are congruences, so is $F \circ G$.*

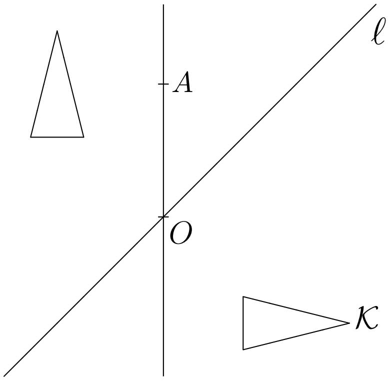
Proof It has been pointed out that every one of the basic isometries has the following three properties: it is a transformation, it is an isometry, and it maps lines to lines as well as preserves the degrees of angles. Because these properties persist under composition, the proof of part (a) of the theorem is straightforward. To prove part (b), i.e., the inverse of a congruence is a congruence, let a congruence φ be the composition of three basic isometries $F \circ G \circ H$, then it is simple to directly verify that if $\psi = H^{-1} \circ G^{-1} \circ F^{-1}$, then $\psi \circ \varphi = I = \varphi \circ \psi$. So ψ is the inverse of φ. But the inverse of a basic transformation is a basic transformation (the inverse of a rotation is a rotation, the inverse of a reflection is itself, and the inverse of a translation is a translation; see §3), so ψ is also a congruence. A similar statement holds if φ is the composition of any number of basic isometries for exactly the same reason. Part (c) follows immediately from the definition of a congruence as a composition of basic isometries. This proves the theorem.

A subset of the plane $\mathcal{S}$ is said to be **congruent to** another subset $\mathcal{S}'$ of the plane if there is a congruence φ so that $\varphi(\mathcal{S}) = \mathcal{S}'$. In symbols: $\boldsymbol{\mathcal{S} \cong \mathcal{S}'}$. Since the inverse of a congruence is a congruence (Theorem G5(b)), and since $\varphi(\mathcal{S}) = \mathcal{S}'$ implies $\varphi^{-1}(\mathcal{S}') = \mathcal{S}$, we see that $\mathcal{S} \cong \mathcal{S}'$ implies $\mathcal{S}' \cong \mathcal{S}$. Thus we can speak unambiguously about **two sets $\boldsymbol{\mathcal{S}}$ and $\boldsymbol{\mathcal{S}'}$ being congruent** since if $\mathcal{S} \cong \mathcal{S}'$, then also $\mathcal{S}' \cong \mathcal{S}$. We leave as an exercise to show that if $\mathcal{S}_1$ is congruent to $\mathcal{S}_2$ and $\mathcal{S}_2$ is congruent to $\mathcal{S}_3$, then $\mathcal{S}_1$ is congruent to $\mathcal{S}_3$. This fact is usually expressed by saying that **congruence is transitive**.

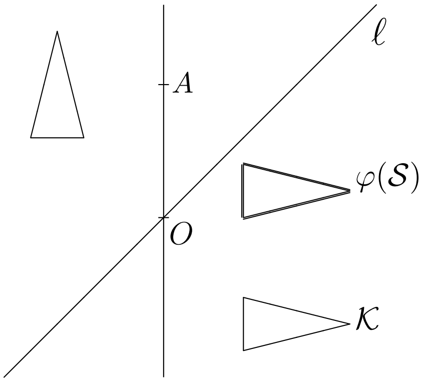
Example Let $\mathcal{S}$ ibe the isosceles triangle shown below. Let R be the reflection across the line ℓ and let T be the translation along the vector $\overline{OA}$, where O is the origin and $A = (0, 5)$. Consider the congruence $\varphi = T \circ R$. What is $\varphi(\mathcal{S})$?



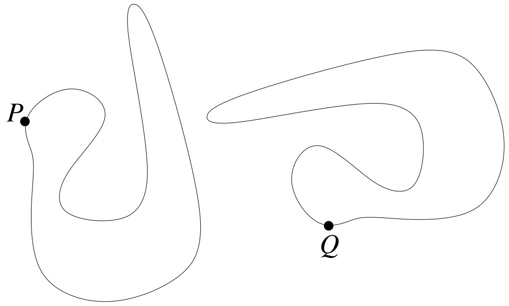
By definition, $\varphi(\mathcal{S}) = T(R(\mathcal{S}))$. If we denote the set $R(\mathcal{S})$ by $\mathcal{K}$, then $\mathcal{K}$ is the triangle in the lower right corner.



Now T moves everything up by the length of OA. The thickened triangle is then $T(\mathcal{K})$, which is $\varphi(\mathcal{S})$.



It is worth pointing out that this definition of congruence applies not only to polygons, but to *any* geometric figures. For example, the following two curves are congruent because, one can map the left curve onto the right curve by a translation along $\overline{PQ}$ followed by a rotation of 90°. *This, and not "same size and same shape", is the meaning that the curves are congruent.*



Congruent triangles occupy a special position in elementary geometry and has its own special conventions. Denote a triangle ABC by $\triangle \boldsymbol{ABC}$. The **congruence notation $\triangle \boldsymbol{ABC} \cong \triangle \boldsymbol{A'B'C'}$** will be understood to mean — in addition to the established meaning that the two sets $\varphi(\triangle ABC)$ and $\triangle A'B'C'$ are equal for some congruence φ — that φ also satisfies $\varphi(A) = A'$, $\varphi(B) = B'$, and $\varphi(C) = C'$.

Congruence criteria

It follows from Theorem G5(a) that, if $\triangle ABC \cong \triangle A'B'C'$, then $\varphi(AB) = A'B'$, $\varphi(AC) = A'C'$, and $\varphi(BC) = B'C'$, and also that $\varphi(\angle A) = \angle A'$, $\varphi(\angle B) = \angle B'$, and $\varphi(\angle C) = \angle C'$. Therefore, again by Theorem G5(a), $\triangle ABC \cong \triangle A'B'C'$ implies that

$$|\angle A| = |\angle A'|, \quad |\angle B| = |\angle B'|, \quad |\angle C| = |\angle C'|,$$

and

$$|AB| = |A'B'|, \quad |AC| = |A'C'|, \quad |BC| = |B'C'|.$$

We now prove the converse. Let it be said that, in this case, the proof of the theorem is as important as the theorem itself.

Theorem G6 *If for two triangles $\triangle ABC$ and $\triangle A'B'C'$,*

$$|\angle A| = |\angle A'|, \quad |\angle B| = |\angle B'|, \quad |\angle C| = |\angle C'|,$$

and

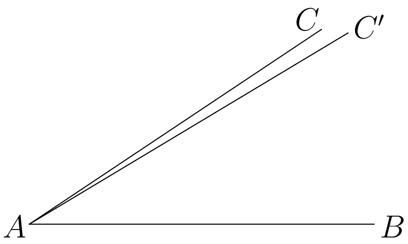
$$|AB| = |A'B'|, \quad |AC| = |A'C'|, \quad |BC| = |B'C'|,$$

then $\triangle ABC \cong \triangle A'B'C'$.

Proof We will prove that if the triangles ABC and $A'B'C'$ satisfy the equalities of sides and angles as described, then there is a congruence φ so that $\varphi(\triangle ABC) = \triangle A'B'C'$. Because we have a precise *definition* of congruence, all we need to do is to find translations, reflections, and rotations that map one triangle to the other. What is worth learning from this proof is the way to break up the rather long process of finding these basic isometries into a sequence of simpler special cases, and of course, how to put everything back together at the end. There are three cases to consider.

Case I: The triangles satisfy the special restriction that $A = A'$, $B = B'$.

In this case, either C, C' are already in the same half-plane of L_{AB}, or they are in opposite half-planes of L_{AB}. If the former, then we claim that $C = C'$, so that in this situation, we need only let φ be I, the identity transformation. To prove the claim, observe that because $|\angle CAB| = |\angle C'AB|$, the fact that C and C' are in the same half-plane of L_{AB} implies that we have the equality of rays, $R_{AC} = R_{AC'}$ (see statement (α) in the discussion of angle measurements in §1).



In like manner, because $|\angle CBA| = |\angle C'BA|$ we have $R_{BC} = R_{BC'}$. Therefore

$$R_{AC} \cap R_{BC} = R_{AC'} \cap R_{BC'}$$

which means of course that $C = C'$. So in this situation, $I(\triangle ABC) = \triangle A'B'C'$.

It remains to deal with the case where $A = A'$, $B = B'$ but C, C' are in opposite half-planes of L_{AB}. Then let R be the reflection with respect to L_{AB}, and let $R(C') = C^*$. Because R maps A' and B' to themselves (they are on L_{AB}), we have $R(\triangle A'B'C') = \triangle A'B'C^*$. Now compare the two triangles ABC and $A'B'C^*$: we have $A = A'$, $B = B'$ as before, but C and C^* are now in the same half-plane of L_{AB}. Because R is a basic isometry, all the expected equalities of sides and angles between the two triangles ABC and $A'B'C^*$ continue to hold. The preceding consideration therefore implies that $I(\triangle ABC) = \triangle A'B'C^*$, and therefore, $I(\triangle ABC) = R(\triangle A'B'C')$. Since $R \circ R = I$, we obtain

$$R(\triangle ABC) = R(I(\triangle ABC)) = R(R(\triangle A'B'C') = \triangle A'B'C'$$

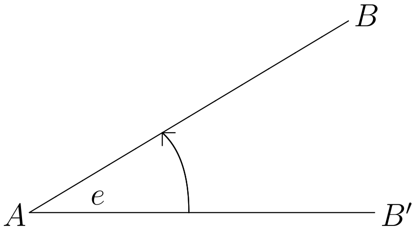
Thus letting φ_1 be either I or R, depending on whether C, C' lie in the same half-plane or different half-planes of L_{AB}, we have

$$\varphi_1(\triangle ABC) = \triangle A'B'C'$$

and Case I is proved.

Case II: The triangles satisfy only the condition that $A = A'$, but nothing about B and B'.

Then there is some number e, $-180 < e \leq 180$, so that the rotation ρ_e around the point A rotates one ray to the other: $\rho_e(R_{AB'}) = R_{AB}$. In the following picture, $e > 0$, but if the points B and B' are interchanged, then we would have to rotate clockwise from B' to B and e would be negative.



Now B and $\rho_e(B')$ (to be called B^*) are two points on the ray R_{AB}, and (because $|AB| = |A'B'|$ and ρ_e is a basic isometry) $|A'B'| = |\rho_e(A'B')| = |AB^*|$. Therefore $B = B^*$, i.e., $B = \rho_e(B')$. Letting $C^* = \rho_e(C')$, we get $\rho_e(\triangle A'B'C') = \triangle ABC^*$. Therefore, the two triangles ABC and ABC^* satisfy the condition of Case I. Consequently, $\varphi_1(\triangle ABC) = ABC^*$ for a basic isometry φ_1. In other words, $\varphi_1(\triangle ABC) =$

$\rho_e(\triangle A'B'C')$. Letting φ_2 be the inverse ρ_{-e} of ρ_e, and applying φ_2 to both sides of this equation, we see that

$$\varphi_2(\varphi_1(\triangle ABC)) = \triangle A'B'C'$$

Thus the theorem is also proved for Case II because $\varphi_2 \circ \varphi_1$ is a congruence.

Case III: The general case where no restriction is placed on the triangles ABC and $A'B'C'$.

We may therefore assume that the vertices A and A' are distinct. Let T be the translation along the vector $\overline{A'A}$; note that $T(A') = A$, so that if we define $B^* = T(B')$ and $C^* = T(C')$, then $T(\triangle A'B'C') = \triangle AB^*C^*$. Now since the triangles ABC and AB^*C^* have the vertex A in common, Case II applies. Thus for suitable basic isometries φ_1 and φ_2, we have $\varphi_2(\varphi_1(\triangle ABC)) = \triangle AB^*C^*$, which is of course equivalent to $\varphi_2(\varphi_1(\triangle ABC)) = T(\triangle A'B'C')$. Let φ_3 be the inverse translation $T_{AA'}$ of T. Then we obtain, for the general case of two triangles ABC and $A'B'C'$ with three pairs of equal sides and equal angles, the fact that

$$\varphi_3(\varphi_2(\varphi_1(\triangle ABC))) = \triangle A'B'C'$$

Letting φ be the congruence $\varphi_3 \circ \varphi_2 \circ \varphi_1$, we see that $\varphi(\triangle ABC) = \triangle A'B'C'$. This completes the proof of Theorem G6.

Theorem G6 is an overkill, in that it is hardly necessary to require the equalities of *all* the angles and *all* the sides of two triangles before we can prove the triangles are congruent. Typically, it suffices to impose three suitably chosen conditions to get it done, and the best known of which are *SAS*, *ASA*, and *SSS*. This is quite clear if one reads the preceding proof carefully and notices that some of the equalities in the hypothesis did not even get mentioned.. Specifically, by retracing the preceding proof, one can see without difficulty how to prove the following two theorems.

Theorem G7 (SAS) *Given two triangles ABC and $A'B'C'$ so that $|\angle A| = |\angle A'|$, $|AB| = |A'B'|$, and $|AC| = |A'C'|$. Then the triangles are congruent.*

Theorem G8 (ASA) *Given two triangles ABC and $A'B'C'$ so that $|AB| = |A'B'|$, $|\angle A| = |\angle A'|$, and $|\angle B| = |\angle B'|$. Then the triangles are congruent.*

Consider the proof of Theorem G8, for example. If we examine the proof of Theorem G6 closely, we will see that all we ever use are the assumptions that

$$|AB| = |A'B'|, \quad |\angle A| = |\angle A'|, \quad \text{and} \quad |\angle B| = |\angle B'|$$

This is because in Case 1, these are the only facts we need to guarantee that $C = C'$ if C and C' are in the same half-plane of L_{AB}, and that the reflection across L_{AB} maps C' to C if C and C' are in opposite half-planes of L_{AB}. Moreover, the proofs of Case II and Case III depend only on the ability to map A' and B' onto A and B and on the truth of Case I. Therefore the remaining assumptions about $|AC| = |A'C'|$, $|BC| = |B'C'|$, and $|\angle C| = |\angle C'|$ are never invoked. This in essence disposes of Theorem G8. The proof of Theorem G7 can be carried out along the same line. We will therefore leave the details to an exercise. There is a third theorem of this genre, SSS, to the effect that two triangles with three pairs of equal sides are congruent. The proof of SSS requires a bit more preparation and will be given in the Appendix.

We have defined the concepts of isometry and congruence, and it is time that we clarify the relationship between the two. As noted in Theorem G5, a congruence preserves the degrees of angles whereas an isometry, by definition, only preserves distance and has nothing to say about degrees of angles. It is a fact, one that will not be proved here, that *every isometry of the plane is a congruence.* In other words, every isometry turns out to be nothing but the composition of a finite number of basic isometries. This underscores the importance of the basic isometries: *they are all you need to know about isometries of the plane.* Moreover, this also implies that if a transformation of the plane preserves distance, then it must be a congruence and therefore it automatically preserves degrees of angles. This is not an obvious fact. You should be careful at this point, however, because *until we can prove this fact, we cannot assume that an isometry preserves degrees of angles.*

Exercises 5.4

1. Prove that congruence is transitive.

2. Let $ABCD$ be a parallelogram. Then $\triangle ABD \cong CDB$.

3. Let M be the midpoint of a segment AB and let D be a point on the perpendicular bisector of AB. Show that the ray R_{DM} is the angle bisector of $\angle ADB$.

4. Using the notation and picture of the Example in this section, exhibit $R(T(\mathcal{S}))$. Is it equal to $T(R(\mathcal{S}))$?

5. Any two circles of the same radius are congruent. (*Caution*: This is a slippery proof. Be very precise.)

6. (a) Prove Theorem G7. (b) Prove Theorem G8.

7. Two rectangles with two pairs of equal sides are congruent, i.e., there is a congruence that maps one to the other. (*Caution:* Be very careful with what you write down.)

8. The angle bisector from a vertex of a triangle is perpendicular to the opposite side if and only if the two sides of the triangle issuing from this vertex are equal. (Note that by the Crossbar Axiom, there is no question that the angle bisector must intersect the opposite side.)

9. Let $ABCD$ be a parallelogram. If a diagonal is an angle bisector (e.g., BD bisects $\angle ABC$), then all four sides of $ABCD$ are equal. (*Note:* Such a parallelogram will turn out to be a square, but we won't be able to prove that until we know a few more things about the angle sum of a triangle.)

10. Explain why two triangles with two pairs of congruent sides and one pair of congruent angles need not be congruent. (We note that, on the other hand, two triangles with two pairs of equal angles and a pair of equal sides must be congruent. This is so because one can prove that the sum of (the degrees of) angles in a triangle is 180 degrees; therefore, the third pair of angles of these two triangles must be equal as well, and Theorem G8 applies.)

5 A brief pedagogical discussion

In a middle school classroom, if congruence and dilation (which will be formally taken up in the Chapter 6) are going to be taught at all, they should be taught in a way that is closer to Chapter 4 than to Chapter 5 and Chapter 6.[60] For example, the *idea* of the proof of Theorem G6 can be conveyed much more easily by moving concrete models of identical triangles across the blackboard than by the rather formidable symbolic notations of the above proof. Another example is the use of transparencies in Chapter 4 to get across the concepts of the basic isometries: it is much better suited to a middle school classroom than the precise definitions of this chapter. Therefore, for the immediate needs of teaching middle school geometry, Chapter 4 is a better model than Chapters 5 and 6, and it is strongly suggest that you plan your lessons according to Chapter 4, at least to the extent possible. It is much more important to build up students' geometric intuition than overwhelm them with technicalities.

For the purpose of *professional development*, however, Chapters 5 and 6 are indispensable. Chapter 4 is an intentionally oversimplified version of geometry; the gain of wider accessibility is achieved at the price of possible mathematical distortions. If you do not know the mathematics in Chapters 5 and 6, you will be misled into thinking that the oversimplifications in Chapter 4 are the norm, and soon you will be teaching a distorted version of the basic isometries. You may have noticed that the discussion in Chapter 4 begins with translations and ends with rotations, whereas the order is reversed in this chapter. That is no accident. Translations are conceptually simpler than rotations, but it so happens that pinning down with mathematical precision the concept of "direction" (which is needed for the definition of translations) requires more work than defining a circle (in the definition of a rotation). This is why in a *mathematically correct* presentation of the basic isometries, translation has to come last, and a teacher should be well aware of that.

There is a broader reason why we believe that you as a teachers should know the materials of Chapters 5 and 6 before you teach the materials of Chapter 4. The fact, not generally recognized, is that it takes more knowledge to simplify and distill the

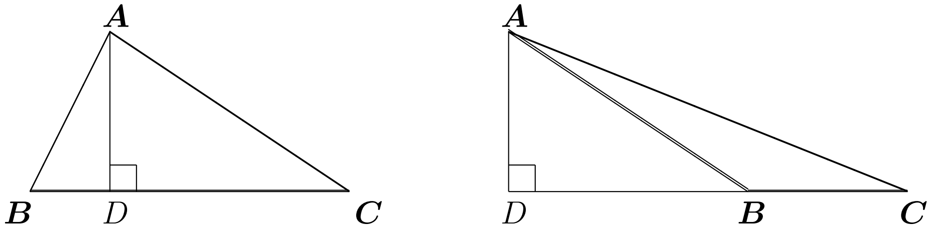
[60] *Added October, 2011:* The Common Core Standards have now made this approach to school geometry—specifically the geometry of grade 8 and high school—a reality in the schools. The geometry curriculum in middle and high schools, as described in these standards, is very close to the content of Chapters 4–6. This gives fresh incentive to the learning of this material.

essence of something than to tell it verbatim. Thus, if you are going to take the challenge of writing a three-page synopsis of *War and Peace*, you'll need a thorough knowledge of the whole book and an understanding of Tolstoy's intentions in order to compress in the short space of three pages what Tolstoy had the luxury of more than a thousand pages to convey. Every word must count. So it is that when you try to give middle school students a taste of geometry, every word must count, and you had better get it right from the start. Reading and understanding Chapters 5 and 6 will give you a better perspective on these geometric facts and concepts. They will decrease the likelihood that you will mislead your students.

Appendix

Our goal is to prove the standard theorem that two triangles whose three pairs of corresponding sides are equal must be congruent (SSS). The three theorems, SAS (Theorem G7), ASA (Theorems G8), and now SSS, form the cornerstone of the discussion of triangles in the high school geometry curriculum. The proof of SSS requires some simple properties of isosceles triangles.

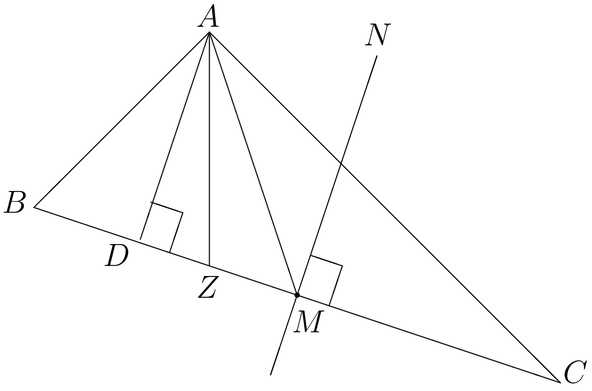
First, some definitions. Recall that a triangle is *isosceles* if it has (at least) two equal sides. If in $\triangle ABC$, $|AB| = |AC|$, we will, for convenience, refer to A as the **top vertex** and $\angle A$ as the **top angle**. The angles $\angle B$ and $\angle C$ are the **base angles**, and side BC is the **base**. Recall also that, for a general triangle, the *line* passing through A and perpendicular to BC is called **the altitude on $\boldsymbol{BC}$** or sometimes, the **altitude from $\boldsymbol{A}$**. Sometimes, the *segment* AD, where D is the point of intersection of this perpendicular line with BC, is also called **the altitude from $\boldsymbol{A}$**. This is a typical example of the same term being used in school mathematics for two different purposes.



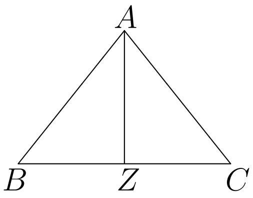
There is no such ambiguity about the next definition, however. The segment joining a vertex to the midpoint of the opposite side is by definition a **median** of the triangle.

Lemma (a) *Isosceles triangles have equal base angles.* (b) *In an isosceles triangle, the perpendicular bisector of the base, the angle bisector of the top angle, the median from the top vertex, and the altitude on the base all coincide.*

To put assertion (b) of the Lemma in context, observe that in a general triangle, the following four lines are distinct: the angle bisector AZ of angle $\angle A$, the median AM and the altitude AD on BC, and the perpendicular bisector NM of the side BC opposite A. See the picture below. It is therefore remarkable that in the case of an isosceles triangle, all four lines would collapse into one at least for the base and the top angle.



Proof Let $|AB| = |AC|$ in $\triangle ABC$, and let the angle bisector of the top angle $\angle A$ intersect the base BC at Z.[61] Let R be the reflection across the line L_{AZ}.



Since $|\angle BAZ| = |\angle CAZ|$, and since reflections preserves the degrees of angles, R maps the ray R_{AB} to the ray R_{AC}. Let B' be the point on the ray R_{AC} so that $R(B) =$

[61] The fact that the angle bisector of $\angle A$ intersects BC is implied by the Crossbar Axiom.

B'. Since A is on the line of reflection, $R(A) = A$ so that $|AB| = |R(AB)| = |AB'|$. By hypothesis, $|AB| = |AC|$, and therefore$|AC| = |AB'|$. Since C and B' are both on the same ray R_{AC}. we have $R(B) = C$. Now it is also true that $R(Z) = Z$ and $R(A) = A$ because Z and A lie on the line of reflection of R, so $R(BZ) = CZ$ and $R(BA) = CA$ (a reflection maps a segment to a segment). Therefore $R(\angle B) = \angle C$. Since a reflection preserves the degree of angles, $|\angle B| = |\angle C|$. This proves part (a). For part (b), observe that since L_{AZ} is the line of reflection and $R(B) = C$,

$$R(\angle AZB) = \angle AZC \quad \text{and} \quad R(BZ) = (CZ)$$

Therefore $|\angle AZB| = |\angle AZC| = 90°$, and $|BZ| = |CZ|$, so that L_{AZ} is the perpendicular bisector of BC. Since L_{AZ} is, by construction, also the angle bisector of $\angle A$, every statement in (b) follows. The proof is complete.

As an application of the Lemma, we can now prove the SSS criterion of triangle congruence.

Theorem (SSS) *Given triangles ABC and $A'B'C'$. If the three sides are pairwise equal, i.e., $|AB| = |A'B'|$, $|AC| = |A'C'|$, and $|BC| = |B'C'|$, then the triangles are congruent.*

Proof As in the case of the SAS and ASA theorems, we break up the proof into three steps, going from a special case to the most general.

Case I. The two triangles satisfy in addition, $A = A'$ and $B = B'$.

Case II. The triangles satisfy in addition, $A = A'$.

Case III. The general case.

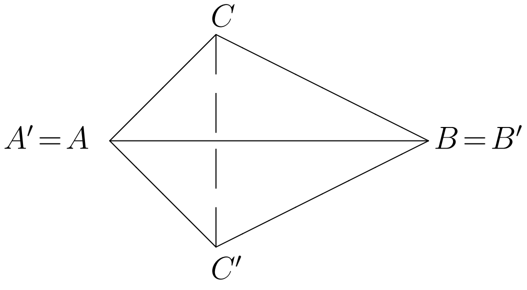
Case I. In this case, there are two possibilities.

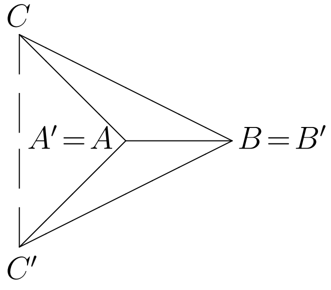
(1) C and C' are in opposite half-planes of L_{AB}.

(2) C and C' are in the same half-plane of L_{AB}.

If (1) holds, then the fact that $|AC| = |A'C'|$ implies that $|\angle ACC'| = |\angle AC'C|$ (see Lemma). Likewise, the fact that $|BC| = |B'C'|$ implies that $|\angle BCC'| = |\angle BC'C|$.

Now there are three possibilities for the two triangles; two are shown below, and the third one is similar to the picture on the right but with AC (respectively, $A'C'$) as the longest side of $\triangle ABC$ (respectively, $\triangle A'B'C'$).





We claim that $|\angle ACB| = |\angle AC'B|$. Indeed, in the situation of the left picture, we have

$$|\angle ACB| \;=\; |\angle ACC'| + |\angle BCC'| \;=\; |\angle AC'C| + |\angle BC'C| \;=\; |\angle AC'B|,$$

whereas in the situation of the right picture,

$$|\angle ACB| \;=\; |\angle BCC'| - |\angle ACC'| \;=\; |\angle BC'C| + |\angle AC'C| \;=\; |\angle AC'B|.$$

It is clear that the third possibility alluded to above can be handled exactly as in the picture on the right. The claim is proved. Thus $\triangle ACB \cong \triangle AC'B$ by SAS (Theorem G7). Consequently, $|\angle CAB| = |\angle C'AB|$ and AB is the angle bisector of the isosceles triangle $AC'C$ and hence the perpendicular bisector of segment CC', by the Lemma again.

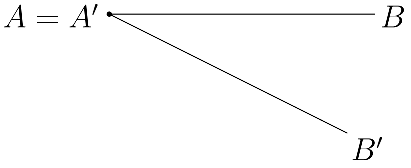
Let R be the reflection across L_{AB}, then by the definition of a reflection, $R(C') = C$ while $R(A') = A$ and $R(B') = B$. Therefore $R(\triangle A'B'C') = \triangle ABC$, and the triangles are congruent, as desired.

If (2) holds instead, then we claim $C = C'$, so that $\triangle ABC = \triangle A'B'C'$. To prove the claim, again let R be the reflection across L_{AB}. Let $R(A') = A_0$, $R(B') = B_0$, and $R(C') = C_0$. Of course, $A_0 = A$ and $B_0 = B$, but now C_0 and C are on opposite half-planes of L_{AB}. The preceding discussion of possibility (1) implies that $R(\triangle A_0B_0C_0) = \triangle ABC$. But $R \circ R = \text{identity}$, so

$$\triangle ABC = R(\triangle A_0B_0C_0) = R(\,R(\triangle A'B'C')\,) = \triangle A'B'C'\,,$$

exactly as claimed, and Case I is completely proved.

Case II. the two triangles have one pair of corresponding vertices in common, say, $A = A'$.



Let ϱ be the rotation around A that brings the ray $R_{AB'}$ to the ray R_{AB}. Because $|AB| = |A'B'|$ by hypothesis, $\varrho(B') = B$. Therefore if we denote $\varrho(A')$, $\varrho(B')$ and $\varrho(C')$ by A_0, B_0 and C_0, respectively, then the triangle ABC and $A_0B_0C_0$ have two pairs of corresponding vertices in common, namely, $A = A_0$ and $B = B_0$. By Case I, there is a congruence φ so that $\varphi(\triangle A_0B_0C_0) = \triangle ABC$. Hence

$$\triangle ABC = \varphi(\triangle A_0B_0C_0) = \varphi(\,\varrho(\triangle A'B'C')\,) = (\varphi \circ \varrho)(\triangle A'B'C')$$

This shows the triangles ABC and $A'B'C'$ are congruent.

Case III. Finally, we deal with the general case. Let T be the translation that maps A' to A, and denote $T(A')$, $T(B')$ and $T(C')$ by A_0, B_0, and C_0, respectively. Then $T(A') = A$ and $\triangle A_0B_0C_0 = T(\triangle A'B'C')$. But $\triangle A_0B_0C_0$ and $\triangle ABC$ now have a pair of corresponding vertices in common because $A_0 = A$. By Case II, there is a congruence φ so that $\varphi(\triangle A_0B_0C_0) = \triangle ABC$. Therefore

$$\triangle ABC = \varphi(\triangle A_0B_0C_0) = \varphi(\,T(\triangle A'B'C')\,) = (\varphi \circ T)(\triangle A'B'C')$$

Hence we have proved the congruence of the triangles ABC and $A'B'C'$ in general when they have three pairs of equal sides. The proof of Theorem G26 is complete.

As an immediate application of SSS, we have the following two corollaries. Their proofs are simple enough to be left as exercises.

Corollary 1 *A quadrilateral is a parallelogram* $\iff$ *its opposite sides are equal.*

Corollary 2 *A rhombus is a parallelogram.*

Chapter 6: Dilation and Similarity

This chapter introduces the other basic concept in school geometry: similarity. Like congruence, similarity has not fared well in school mathematics. Middle school students are taught that two geometric figures are similar if they are *the same shape but not the same size*. Then when they get to high school, they are told that similarity means equal angles and proportional sides. In other words, suddenly "similarity" becomes synonymous with the "similarity of triangles". What such instructions leave behind is a vacuum about what it means for two geometric shapes, which are not polygons (e.g., two ellipses), to be similar. Consequently, there is a high probability that, upon graduation from high school, students' understanding of similarity consists of two disconnected sound bites: a definition of similar triangles in terms of proportional sides and equal angles, and a conception of "same shape but not the same size" for anything other than triangles. What seems to have passed unnoticed is the fact that a correct description of similarity, one that is discussed below, can be easily introduced in middle school through ample hands-on experiments plus a judicious amount of reasoning. In any case, we have to get rid of "same shape but not same size" in the school classroom.

A quick perusal of §5 of the last chapter may lend some perspective on the material of this chapter.

1 The Fundamental Theorem of Similarity

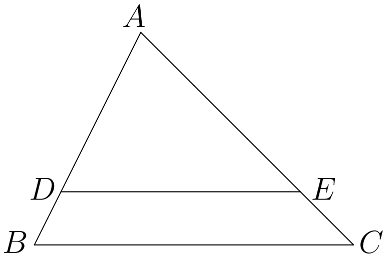
The purpose of this section is to give a *partial* explanation of the following **Fundamental Theorem of Similarity (FTS)**. This theorem dominates the whole discussion of similarity.

Theorem G9 (FTS) *Let $\triangle ABC$ be given, and let D, E be points on AB and AC respectively. If*

$$\frac{|AB|}{|AD|} = \frac{|AC|}{|AE|}$$

(*and their common value is denoted by r*), *then $DE \parallel BC$ and*

$$\frac{|BC|}{|DE|} = r$$



First of all, a remark about terminology. The number r is generally referred to as the **scale factor**. The statement above on $\boldsymbol{DE} \parallel \boldsymbol{BC}$ is a *standard abuse of notation for* $L_{DE} \parallel L_{BC}$, i.e., the line containing the segment DE is parallel to the line containing the segment BC. We will continue to use this abuse of notation for the rest of this book.

In applications, it is sometimes more convenient to assume, instead of $\frac{|AB|}{|AD|} = \frac{|AC|}{|AE|}$, the equivalent condition

$$\frac{|AD|}{|DB|} = \frac{|AE|}{|EC|}$$

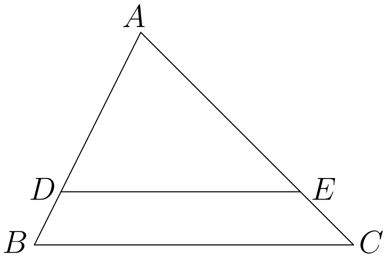
See problem 1 of Exercises 5.1.

In this section, will we give a proof of this theorem for the special case of $r = 2$, i.e., for the case that D and E are midpoints of AB and AC, respectively. The complete proof of FTS for all fractional values of r can be given now, and the only reason for not giving it is that the rather intricate reasoning would be too much of a distraction at this point. Note that, as usual, we have to appeal to FASM for the validity of FTS for *all* positive numbers r rather than just fractional values.

We should point out a different formulation of FTS which is sometimes more convenient for applications:

Theorem G10 (FTS*) *Let $\triangle ABC$ be given, and let $D \in AB$. Suppose a line parallel to BC and passing through D intersects AC at E. Then*

$$\frac{|AB|}{|AD|} = \frac{|AC|}{|AE|} = \frac{|BC|}{|DE|}$$



The simple proof that, because of the presence of the Parallel Postulate, Theorems G10 follows from Theorem G9 and that, conversely, Theorem G9 follows from Theorem G10 will be left as an exercise.

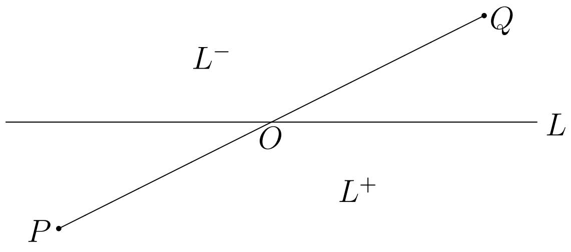
The proof of FTS for the special case of $r = 2$ requires some preparation in the form of the following procession of theorems. Notice the striking fact that every single one of them relies on the fact that a rotation of 180 degrees preserves distance and degrees of angles.

Theorem G11 *Let O be a point on a line L, and let ϱ be the rotation of* 180° *around O. Then ϱ maps each half-plane of L to its opposite half-plane.*

Proof Let the half-planes of L be L^+ and L^-. The theorem says

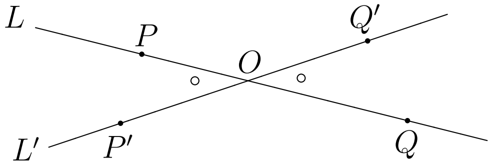
$$\varrho(L^+) = L^- \quad \text{and} \quad \varrho(L^-) = L^+$$

It suffices to prove the first assertion, i.e., $\varrho(L^+) = L^-$. Let us first prove that $\varrho(L^+) \subset L^-$. So let P be a point in L^+, and we will prove $\varrho(P) \in L^-$. On the line L_{PO} joining P to O, let Q be the point on the other half-line of L_{PO} relative to O so that the distance from O to P is equal to the distance from O to Q. It follows that $Q = \varrho(P)$.



Now the segment PQ contains a point of L, namely, O, so P and Q are in opposite half-planes of L (see (L4) in §1). Since $P \in L^+$, we have $Q \in L^-$, i.e., $\varrho(P) \in L^-$, as claimed. Next we need to prove that $L^- \subset \varrho(L^+)$. Thus given $Q \in L^-$, we must show that there is a point $P \in L^+$ so that $\varrho(P) = Q$. We reverse the preceding argument: Join Q to O to obtain the line L_{QO}, and on this line, take the point P on the other half-line relative to O so that P, Q are equi-distant from O. Then by definition, $\varrho(P) = Q$. Since the segment PQ contains a point of L (namely, O), P and Q have to be in opposite half-planes. Thus $P \in L^+$, and the proof is complete.

Let L and L' be two lines meeting at a point O. On L (resp., L'), let P, Q (resp., P', Q') be points lying on opposite half-lines determined by O, as shown in the following figure.



Then the angles $\angle POP'$ and $\angle QOQ'$ are called **opposite angles**.[62] We have:

Theorem G12 *Opposite angles are equal.*

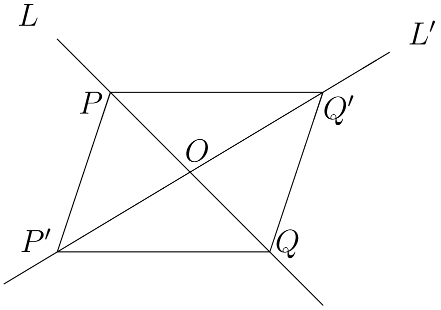
Proof We make use of the preceding figure. The proof is rather trivial: each of the two numbers, $|\angle POP'|$ and $|\angle QOQ'|$, when added to $|\angle P'OQ|$ is 180 because $\angle POQ$ and $\angle P'OQ'$ are both straight angles. So $|\angle POP'| = |\angle QOQ'|$. We want to give a different proof, however, because our purpose is to demonstrate how to make use of basic isometries to prove theorems. In this case, we argue as follows. Consider the rotation ϱ of 180° around O. Clearly $\varrho(R_{OP}) = R_{OQ}$ and $\varrho(R_{OP'}) = R_{OQ'}$. Therefore $\varrho(\angle POP') = \angle QOQ'$. Since ϱ preserves angles (by assumption (ϱ3) of rotations), we have $|\angle POP'| = |\angle QOQ'|$.

The next two theorems give characterizations of a parallelogram that will prove to be useful. The first one says that parallelograms are the quadrilaterals whose di-

[62] Most school textbooks in the U.S. call these *vertical angles*.

agonals bisect each other.

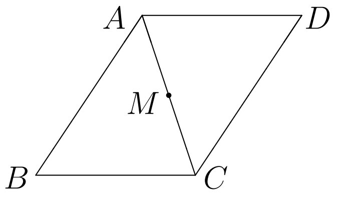
Theorem G13 *Let L and L' be two lines meeting at a point O. P, Q (resp., P', Q') are points lying on opposite half-lines of L (resp., L') determined by O. Then $|PO| = |OQ|$ and $|P'O| = |OQ'| \iff PP'QQ'$ is a parallelogram.*



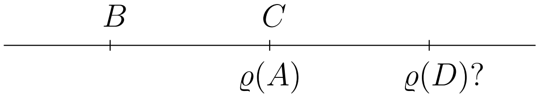
Proof We will prove that if $|PO| = |OQ|$ and $|P'O| = |OQ'|$, then $PP'QQ'$ is a parallelogram; the converse will be left as an exercise. As usual, let ϱ be the rotation of 180° around O. Then ϱ interchanges the rays of L_{PQ} with vertex at O. Thus $\varrho(R_{OP}) = R_{OQ}$, so that $\varrho(P) \in R_{OQ}$. But ϱ is an isometry, so $|\varrho(OP)| = |OP|$, or written differently, $|O\varrho(P)| = |OP|$. By hypothesis, $|OP| = |OQ|$, so $|O\varrho(P)| = |OQ|$. Since both $\varrho(P)$ and Q are in R_{OQ}, we conclude that $\varrho(P) = Q$. Similarly, $\varrho(P') = Q'$, so that $\varrho(PP') = QQ'$. By Theorem G1, $PP' \parallel QQ'$. In the same way, we can prove $PQ' \parallel P'Q$. This proves that $PP'QQ'$ is a parallelogram, as desired.

Theorem G14 *A quadrilateral is a parallelogram $\iff$ it has one pair of sides which are equal and parallel.*

Proof The fact that a parallelogram has a pair of sides which are equal and parallel is implied by Theorem G4. We prove the converse. Let $ABCD$ be a quadrilateral so that $|AD| = |BC|$ and $AD \parallel BC$. We have to prove that $ABCD$ is a parallelogram. It suffices to prove that $AB \parallel CD$. Let ϱ be the rotation of 180 degrees around the midpoint M of the diagonal AC.



As usual, $\varrho(A) = C$ by the definition of ϱ, and $\varrho(L_{AD}) \parallel L_{AD}$, by Theorem G1. Therefore $\varrho(L_{AD})$ is a line passing through C and parallel to L_{AD} itself. Since L_{BC} is also a line passing through C and parallel to L_{AD}, the Parallel Postulate implies that $\varrho(L_{AD}) = L_{BC}$. At this juncture, we only know that $\varrho(D)$ lies in L_{BC}, but we are going to show that, in fact, $\varrho(D) = B$. To this end, observe that on the line L_{BC}, there are now three points: B, $\varrho(D)$, and C, and of course C coincides with $\varrho(A)$. We want to show that $\varrho(D)$ also coincides with B. Since ϱ is an isometry, $|\varrho(AD)| = |AD|$. But $|AD| = |BC|$ by hypothesis, so $|\varrho(AD)| = |BC|$. Since ϱ maps a segment to a segment, this equality says ϱ maps AD to a segment in L_{BC} of length $|BC|$ joining $\varrho(A)$ (which is just C) to $\varrho(D)$.

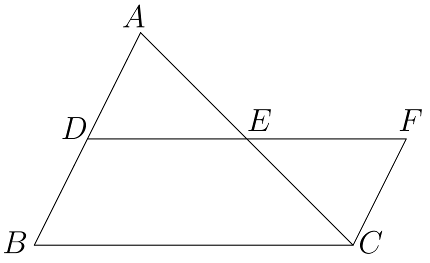


But BC is also a segment of length $|BC|$ that has C as an endpoint, so to prove $\varrho(D) = B$, we have to prove that on the line L_{BC}, both B and $\varrho(D)$ lie in the same half-line of line L_{BC} with respect to C. By Theorem G11, D and $\varrho(D)$ must lie in opposite half-planes of L_{AC}. Since D and B also lie in opposite half-planes of L_{AC}, B and $\varrho(D)$ must lie in the same half-plane of L_{AC}, and therewith, also on the same ray R_{CB} (for the reason that the intersection of the closed half-plane of L_{AC} containing B with the line L_{BC} is the ray R_{CB}). Hence B and $\varrho(D)$, being two points of the same distance from C on the same ray R_{CB}, must coincide, i.e., $\varrho(D) = B$. Coupled with the fact that $\varrho(C) = A$, we see that $\varrho(CD) = AB$, and therefore, $\varrho(L_{CD}) = L_{AB}$. But according to Theorem G1, $\varrho(L_{CD}) \parallel L_{CD}$. Hence $L_{AB} \parallel L_{CD}$, as desired.

Remark In the preceding proof, the fact that B and D lie in opposite half-planes of the diagonal line L_{AC} was taken for granted. This assumption allowed us to conclude that $B = \varrho(D)$. This assumption was made in the proof because the proof of this intuitively obvious fact is too intricate to be of interest in a middle school classroom.

We are finally in a position to prove the special case of FTS when $r = 2$:

Theorem G15 *Let $\triangle ABC$ be given, and let D and E be midpoints of AB and AC, respectively. Then $DE \parallel BC$ and $|BC| = 2|DE|$.*



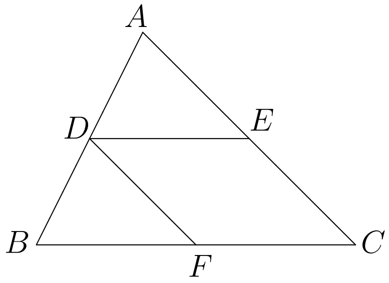
Proof It is not obvious, but the key idea is to look for a point F on the ray R_{DE} so that $|DE| = |EF|$. (Once you buy into this idea, you can do variations on this theme and come up with different proofs.)

Consider the rotation ϱ of 180 degrees around E. Since A and C are equi-distant from E, as are D and F, we have $\varrho(CF) = AD$. Since ϱ is an isometry, $|CF| = |AD|$, but since $|AD| = |DB|$ by hypothesis, we have $|CF| = |DB|$. On the other hand, by Theorem G1 of Chapter 4, §3, $CF \parallel AD$, which is of course the same as $CF \parallel BD$. The quadrilateral $DBCF$ therefore has a pair of sides which are equal and parallel. By Theorem G14, $DBCF$ is a parallelogram. Thus $DF \parallel BC$, which is the same as $DE \parallel BC$. Furthermore, $|DF| = |BC|$ (Theorem G4), and since $|DE| = |EF|$, we have $|BC| = 2|DE|$. The proof is complete.

Theorem G15 has a surprising consequence: if $ABCD$ is any quadrilateral, then the quadrilateral obtained by joining midpoints of the adjacent sides of $ABCD$ is always a parallelogram. (See problem 4 in Exercises 5.1 immediately following).

Because of the importance of Theorem G15 in our work, we will give a second proof using translations. The strategy is to first prove a special case of Theorem G10 (FTS*), and then use it to get at Theorem G15. So our first task is to prove the following theorem.

Theorem G15* *Let $\triangle ABC$ be given and let D be the midpoint of AB. Suppose a line parallel to BC passing through D intersects AC at E. Then E is the midpoint of AC and* $2|DE| = |BC|$.



Proof Let T denote the translation along the vector $\overline{AD}$. Because $|AD| = |DB|$ by hypothesis, the definition of T implies that $T(D) = B$. From §3 of Chapter 4, we also know that T maps any line L not parallel to AD to another line parallel to L itself. Therefore $T(L_{DE})$ is a line passing through B and parallel to DE. By hypothesis again, we already know $L_{BC} \parallel DE$. By the Parallel Postulate, we see that $T(L_{DE}) = L_{BC}$. In particular, $T(E)$ is a point F on BC, i.e.,

$$T(E) \;=\; F$$

Therefore we have $T(DE) = BF$. Because T is an isometry, we have also

$$|DE| \;=\; |BF|$$

Now consider $T(L_{AC})$. Because $T(A) = D$ and $T(E) = F$, it follows that $T(AE) = DF$ and $T(L_{AC}) \;=\; L_{DF}$. Using once more the fact that a translation is an isometry that maps a line to another line parallel to itself, we have

$$DF \parallel AC, \quad \text{and} \quad |AE| = |DF|$$

Since $DE \parallel BC$ by hypothesis, $DFCE$ is a parallelogram and therefore,

$$|DE| \;=\; |FC| \quad \text{and} \quad |DF| \;=\; |EC|$$

by Theorem G4. Since $|DE| = |BF|$, the first equality implies $2|DE| = |BC|$. Since also $|AE| = |DF|$, the second equality implies E is the midpoint of AC. The proof of Theorem G15* is complete.

Now we prove Theorem G15 again. Using the notation and picture of that theorem, we draw a line L through D parallel to BC. By what we have just proved, L passes through the midpoint E of AC and therefore $DE \parallel BC$. Also we already know that $2|DE| = |BC|$. We are done.

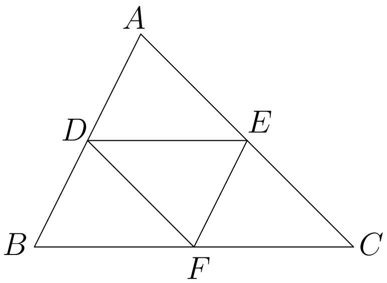
Exercises 6.1

1. Let D and E be points on sides AB and AC, respectively, of $\triangle ABC$. Prove:

$$\frac{|AB|}{|AD|} = \frac{|AC|}{|AE|} \quad \Longleftrightarrow \quad \frac{|AD|}{|DB|} = \frac{|AE|}{|EC|} \Longleftrightarrow \quad \frac{|AB|}{|DB|} = \frac{|AC|}{|EC|}$$

2. Finish the proof of Theorem G13 by proving that the diagonals of a parallelogram **bisect each other**, in the sense that their point of intersection is the midpoint of each diagonal. (*Caution:* Be careful.)

3. Let D, E, F be the midpoints of sides AD, AC, and BC, respectively, of $\triangle ABC$. Then the four triangles ADE, DBF, DEF, EFC are all congruent.



4. If $ABCD$ is any quadrilateral, then the quadrilateral obtained by joining midpoints of the adjacent sides of $ABCD$ is always a parallelogram.

5. Let L_1, L_2, and L_3 be three mutually parallel lines, and let ℓ and ℓ' be two distinct transversals. If the two segments intercepted on ℓ by L_1, L_2, and by L_2, L_3 are of the same length, then the same is true of the corresponding segments intercepted on ℓ'. (There are many ways to prove this, but the most natural is to make use of Theorem G15*.)

6. Use the idea in the proof of Theorem G15, *but do not assume FTS*, to prove that if in triangle ABC, D and E are points on AB and AC respectively, so that $|AB| = 3|AD|$ and $|AC| = 3|AE|$, then $DE \parallel BC$ and $|BC| = 3|DE|$.

7. (a) *FTS* (*Theorem G9*) *implies FTS** (*Theorem G10*). More precisely, this means: assume everything we have proved up to and including Theorem G9 and prove Theorem G10. (b) *FTS** (*Theorem G10*) *implies FTS* (*Theorem G9*). More precisely, this means: assume everything we have proved up to and including Theorem G8, plus Theorem G10, and prove Theorem G9. (*The standard mathematical statement that succinctly summarizes* (*a*) *and* (*b*) *is that* "Theorem G9 and Theorem G10 are equivalent".)

8. Given positive numbers a and b, prove that there exists a rectangle whose sides have lengths a and b. (*Don't skip any steps!*)

9. Let F be the midpoint of the side BC of $\triangle ABC$. Then AF is the angle bisector of $\angle A$ if and only if AB and AC are equal.

10. Let $ABCD$ be a parallelogram. Then B, D lie in opposite half-planes of the diagonal AC, or more correctly, lie on opposite half-planes of the line L_{AC}.

11. Let $A = (a, a')$ and $B = (b, b')$. Prove that the midpoint of the segment AB is $(\frac{a+b}{2}, \frac{a'+b'}{2})$.

2 Dilation: Part II

We have been considering isometries almost exclusively thus far. Now we have to look seriously into an important class of transformations that are not isometries.

Definition *A transformation D of the plane is* **a dilation with center $\boldsymbol{O}$ and scale factor $\boldsymbol{r}$ ($\boldsymbol{r > 0}$)** *if*

(1) $D(O) = O$.

(2) *If $P \neq O$, the point $D(P)$, to be denoted simply by Q, is the point on the ray R_{OP} so that $|OQ| = r|OP|$.*

As already noted back in Chapter 4, a dilation with center at O maps each point by "pushing out" or "pulling in" the point along the ray from O to that point, depending on whether the scale factor r is bigger than 1 or smaller than 1. In particular, each ray issuing from O is mapped into itself.

We hasten to confirm that a dilation so defined *is* a transformation as claimed, i.e., that D satisfies the two defining properties of a transformation:

(i) if P_1 and P_2 are distinct points of Π, then D assigns to them distinct points $D(P_1)$ and $D(P_2)$ of Π, and

(ii) for every point Q of Π there is a point P of Π so that D assigns P to Q, i.e., $Q = D(P)$ for some point P.

To see that (i) is true, first observe that if one of the two distinct points P_1 and P_2 is equal to O, let us say $P_1 = O$, then $P_2 \neq O$. By the definition of D, $D(P_2) \neq O$. Since $D(P_1) = D(O) = O$, we see that $D(P_1) \neq D(P_2)$. Thus we may assume that neither P_1 nor P_2 is equal to O. If P_1 and P_2 lie on different rays issuing from O, then by the definition of D, the points $D(P_1)$ and $D(P_2)$ remain on different rays and are therefore not equal to each other. If on the other hand, P_1 and P_2 lie on the same ray issuing from O, let $D(P_1) = P_1'$ and $D(P_2) = P_2'$. Then

$$|OP_1'| \;=\; r\,|OP_1| \;\neq\; r\,|OP_2| \;=\; |OP_2|$$

and hence $P_1' \neq P_2'$, i.e., $D(P_1) \neq D(P_2)$.

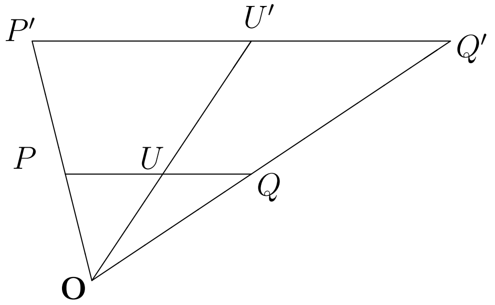
To see that (ii) is true, let Q be given. If $Q = O$, then $D(Q) = Q$. We may therefore assume that $Q \neq O$. On the ray R_{OQ}, let P be chosen so that $|OP| = (1/r)|OQ|$.

Then by the definition of D, $D(P) = Q$. We have therefore proved that a dilation, so defined, is a transformation.

A fundamental property of dilations, one that makes possible the simple drawings of the dilation of rectilinear figures, is the following. It will be clear from this and subsequent proofs related to dilation that the FTS and the Parallel Postulate lie at the heart of the matter.

Theorem G16 *Dilations map segments to segments. More precisely, a dilation D maps a segment PQ to the segment joining $D(P)$ to $D(Q)$. Moreover, if the line L_{PQ} does not pass through the center of the dilation D, then the line L_{PQ} is parallel to the line containing $D(PQ)$.*

Proof Let D have center O and scale factor r. If L_{PQ} passes through O, then either P and Q lie on the same side of O or they lie on opposite sides of O. In either case, the fact that D maps PQ to the segment in L_{PQ} from $D(P)$ to $D(Q)$ follows immediately from the definition of a dilation. We therefore assume that L_{PQ} does not pass through O. Let $P' = D(P)$, $Q' = D(Q)$.



Let U be any point of PQ, and we will show that $D(U)$ is on $P'Q'$. Let $U' = D(U)$. Consider $\triangle OP'U'$. Because D maps P and U to P' and U', respectively,

$$\frac{|OP'|}{|OP|} = \frac{|OU'|}{|OU|} = r$$

By FTS, $L_{PU} \parallel L_{P'U'}$. Denoting L_{PQ} by L, this says $L_{P'U'} \parallel L$. If we apply the same reasoning to $\triangle OP'Q'$, we get

$$\frac{|OP'|}{|OP|} = \frac{|OQ'|}{|OQ|} = r$$

and, therewith, also $L_{P'Q'} \parallel L$. Denoting $L_{P'Q'}$ by L', we therefore have $L' \parallel L$. Thus both L' and $L_{P'U'}$ are lines passing through P' and parallel to L. By the Parallel Postulate, $L_{P'U'} = L'$, and in particular, U' lies on L'. The fact that U' lies in the segment $P'Q'$ follows from the crossbar axiom: the latter implies that, because U lies in $\angle POQ$, the ray R_{OU} must meet the segment $P'Q'$ at some point, which then must be U' because two distinct lines meet only at one point. We pause to observe that we have also proved in the process that $L' \parallel L$, assuming L does not contain O.

Next, we show that $P'Q' \subset D(PQ)$. Let $U' \in P'Q'$. Let U be the intersection of $R_{OU'}$ and L. For exactly the same reason as before, $U \in PQ$. We now prove that $D(U) = U'$. If we can prove that

$$\frac{|OU'|}{|OU|} = r$$

then by the definition of D, we would have $D(U) = U'$. To prove this, let V be the point on R_{OU} so that $|OV| = r|OU|$. In $\triangle OP'V$, we have $|OP'|/|OP| = r$, so that

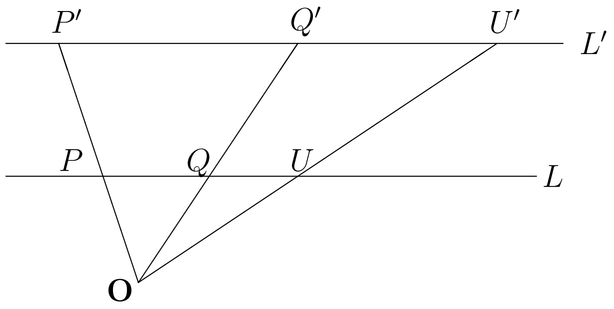
$$\frac{|OP'|}{|OP|} = \frac{|OV|}{|OU|}$$

By FTS, $L_{P'V} \parallel L_{PU}$, or what is the same thing, $L_{P'V} \parallel L$. Since L' is also a line passing through P' and parallel to L, the Parallel Postulate again dictates that $L' = L_{P'V}$. So V in fact lies on L', and as it also lies on the ray R_{OU}, we see that $V = U'$. Thus from $|OV| = r|OU|$, we conclude that $|OU'|/|OU| = r$, as desired. The proof of Theorem G16 is complete.

There are two useful corollaries of this theorem.

Corollary 1 *A dilation maps lines to lines, and rays to rays.*

Proof of Corollary 1 Given a line L, we must prove that $D(L)$ is also a line. Let P, Q be points on L. The theorem says $D(PQ)$ is the segment $P'Q'$, where $P' = D(P)$ and $Q' = D(Q)$. Let L' be the line containing $P'Q'$. Let $U \in L$, we will prove that $D(U) \in L'$. We may assume that U lies outside PQ, in which case, we may assume without loss of generality that $Q \in PU$.



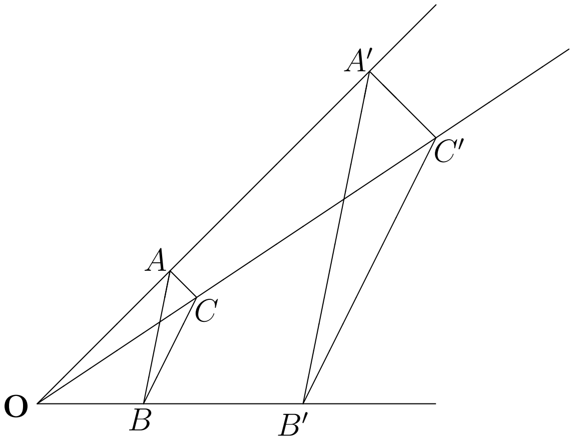
Let $U' = D(U)$. Theorem G16 implies that $D(PU)$ is the segment $P'U'$. Since $Q \in PU$, $Q' \in P'U'$. Thus if L^* denotes the line containing $P'U'$, then P' and Q' both belong to L' and L^*. Therefore $L' = L^*$, and therefore also $U' \in L'$. We have proved that $D(L) \subset L'$. It remains to prove that $L' \subset D(L)$. Let $U' \in L'$, and we let U be the point of intersection of $R_{OU'}$ and L. If we let $V = D(U)$, then we prove exactly as before that $V = U'$, so that $D(L) = L'$.

The fact that D maps rays to rays is proved in the same manner.

Corollary 2 *Let triangles ABC and $A'B'C'$ be given. If a dilation D maps the vertices to vertices, then it also maps the triangle to the triangle. Precisely: if $D(A) = A'$, $D(B) = B'$, and $D(C) = C'$, then $D(\triangle ABC) = \triangle A'B'C'$.*

Proof of Corollary 2 Recall that $\triangle ABC$ is the union of the segments AB, BC, and AC. Thus the conclusion means we must prove $D(AB) = A'B'$, $D(BC) = B'C'$, and $D(AC) = A'C'$. But this follows immediately from Theorem G16 and the hypothesis that $D(A) = A'$, $D(B) = B'$, and $D(C) = C'$.

Armed with Theorem G16, we see that it is very simple to draw the image of a segment by a dilation, a fact that we already discussed in Chapter 4. Indeed, to draw the image of a segment AB by a dilation D, simply find the image points $D(A)$ and $D(B)$ of the endpoints and then draw the segment joining $D(A)$ to $D(B)$. Here is an example with a scale factor of 2.5.



In the picture, $A' = D(A)$, $B' = D(B)$, and $C' = D(C)$.

What about the dilation of *curved* figures? On draws dilated figures through the choice of *data points* in the original figure as we saw in §4 of Chapter 4.

The following theorem summarizes the remaining basic properties of dilations.

Theorem G17 *Let D be a dilation with center O and scale factor r. Then:*

(a) *The inverse transformation of D is the dilation with the same center O but with a scale factor $1/r$.*

(b) *For any segment AB, $|D(AB)| = r|AB|$.*

(c) *D maps angles to angles and preserves degrees of angles.*

REMARK Observe the delicate point that the statements of part (b) and part (c) depend on the validity of Theorem G16 and its Corollary 1. Indeed, without knowing that $D(AB)$ is a segment, the notation $|D(AB)|$ would not even make sense (because the notation $|\sigma|$ only makes sense when σ is a segment or an angle), and without knowing that D maps rays to rays, we would not know that D maps angles to angles.

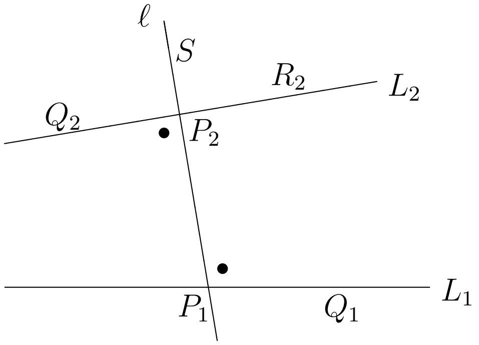
Proofs of parts (a) and (b) (a) Let D' be the dilation with center O and scale factor $\frac{1}{r}$. From the definition of a dilation, it is easy to check that $D \circ D' = I = D' \circ D$. Thus D' is the inverse transformation of D.

Part (b) has been implicitly proved in the proof of Theorem G16. Indeed, in the notation above, if $P' = D(P)$ and $U' = D(U)$, then we have shown that $D(PU) = P'U'$.

If we look at $\triangle OP'U'$, then FTS implies that $|P'U'|/|PU| = r$, i.e., $|D(PU)| = r|PU|$. Since P and U are arbitrary points, (b) is proved.

For the proof of part (c), we need to first get to know something about parallel lines and angles. First some definitions.

Let two distinct lines L_1, L_2 be given. Recall that a *transversal* of L_1 and L_2 is any line ℓ that meets both lines in distinct points. Suppose ℓ meets L_1 and L_2 at P_1 and P_2, respectively. Let Q_1, Q_2 be points on L_1 and L_2, respectively, so that they lie in opposite half-planes of ℓ. Then $\angle Q_1P_1P_2$ and $\angle P_1P_2Q_2$ are said to be **alternate interior angles** of the transversal ℓ with respect to L_1 and L_2.

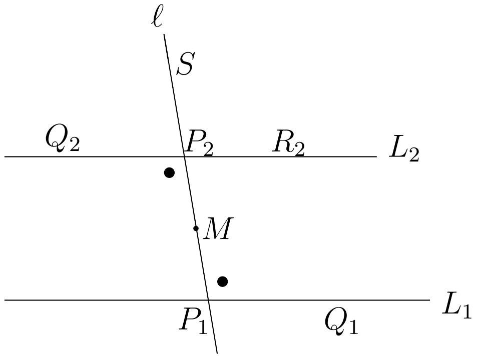


An angle which is the opposite angle of one of a pair of alternate interior angles is said to be the **corresponding angle** of the other angle. For example, because $\angle Q_1P_1P_2$ and $\angle P_1P_2Q_2$ are alternate interior angles and because $\angle SP_2R_2$ in the above figure is the opposite angle of $\angle P_1P_2Q_2$, $\angle SP_2R_2$ is then the corresponding angle of $\angle Q_1P_1P_2$.

In the school classroom, we suggest that alternate interior angles be defined simply by drawing a picture as above and pointing to $\angle Q_1P_1P_2$ and $\angle P_1P_2Q_2$. The correct definition (the one just given), using the precise concept of the half-planes of a line, may be pointed out as a side remark to open students' minds to the potential of complete logical precision. It will be seen presently that we need such precision because we want to present logically complete proofs of several theorems, including the one on the angle sum of a triangle. A overwhelming majority of school students would not take kindly to the need of such precision in the proofs of theorems (as we will do in *the proof of Theorem G18 and later in Chapter 11), because they would consider the investment of so much effort into something so visibly obvious to be ridiculous. So some compromise in the school classroom would be advisable. The purpose of this book is, however, to expand your mathematical horizon for teaching in schools by supplying you with a solid foundation on all things directly related to the K-12 classroom. Acquiring the ability to reason through such bread-and-butter issues as alternate interior angles with precision would certainly serve this purpose admirably.*

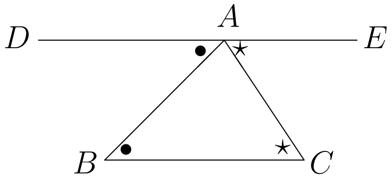
The basic theorem about parallel lines and angles is the following:

Theorem G18 *Alternate interior angles of a transversal with respect to a pair of parallel lines are equal. The same is true of corresponding angles.*



Proof We continue to use the above notation. Let M be the midpoint of P_1P_2, and let ϱ be the rotation of 180 degrees around M. Because $\varrho(P_1) = P_2$, $\varrho(L_1)$ is a line passing through P_2. By Theorem G1, $\varrho(L_1) \parallel L_1$, and the hypothesis says L_2 is also a line passing through P_2 and parallel to L_1, the Parallel Postulate says $\varrho(L_1) = L_2$. By hypothesis, the rays $R_{P_1Q_1}$ and $R_{P_2Q_2}$ lie in opposite half-planes of ℓ, and by Theorem G11, ϱ maps each half-plane of ℓ to its opposite half-plane. Thus $\varrho(R_{P_1Q_1}) = R_{P_2Q_2}$. Of course, $\varrho(R_{P_1P_2}) = R_{P_2P_1}$. Hence, $\varrho(\angle Q_1P_1P_2) = \angle P_1P_2Q_2$, and since ϱ preserves degrees of angles, $|\angle Q_1P_1P_2| = |\angle P_1P_2Q_2|$. The proof is the same for the other pair of alternate interior angles. The last assertion of the theorem about corresponding angles follows from Theorem G12. We have proved Theorem G18.

Remark *Those readers who are familiar with some high school geometry may be very tempted at this point to immediately use Theorem G18 to prove the well-known fact that the sum of (the degrees of) angles in a triangle is* $180°$. *The argument goes as follows. Given* $\triangle ABC$, *draw a line* DE *through* A *that is parallel to* BC, *as shown:*



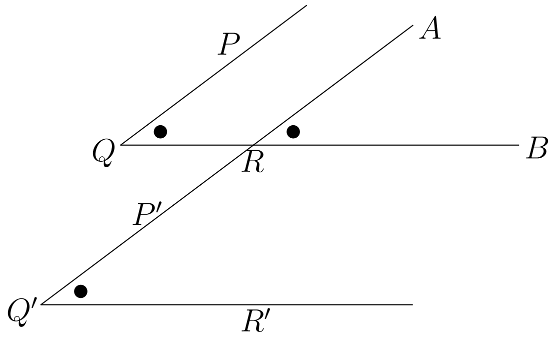
By Theorem G18, $|\angle DAB| = |\angle B|$ *and* $|\angle CAE| = |\angle C|$, *so that*

$$|\angle B| + |\angle BAC| + |\angle C| = |\angle DAB| + |\angle BAC| + |\angle CAE| = 180°$$

This would seem to finish the proof. Let us affirm that this intuitive argument is indeed how a high school student should remember why the angle sum of a triangle is 180. For a teacher *to really come to grips with the delicate points about Euclidean geometry, however, it is necessary to point out that for Theorem G18 to be applicable, we must first prove that* $\angle B$ *and* $\angle DAB$ *are alternate interior angles, as are* $\angle C$ *and* $\angle CAE$. *See the italicized remarks above Theorem G18.*

We can finally finish the **proof of part (c) of Theorem G17**. Since D maps rays to rays, it maps angles to angles. Given $\angle PQR$, let $D(R_{QP}) = R_{Q'P'}$ and $D(R_{QR}) = R_{Q'R'}$, so that $D(\angle PQR) = \angle P'Q'R'$. We have to prove that

$$|\angle PQR| = |\angle P'Q'R'|$$



Without loss of generality, we may assume that both have positive degree. We claim that $L_{Q'P'}$ must intersect L_{QR}. If not, then $L_{Q'P'} \parallel L_{QR}$. But we already know from part (b) that $L_{Q'R'} \parallel L_{QR}$. Thus we have two distinct lines $L_{Q'P'}$ and $L_{Q'R'}$ passing through Q' and parallel to L_{QR}, and this contradicts the Parallel Postulate. Thus $L_{Q'P'}$ intersects L_{QR}. In the interest of notational economy, let the point of intersection continue to be denoted by R, as shown. By Theorem G16, $L_{QR} \parallel L_{Q'R'}$ and $L_{QP} \parallel L_{Q'P'}$. Therefore, according to Theorem G18 about corresponding angles, (notation as in the preceding figure) $|\angle PQR| = |\angle ARB| = |\angle P'Q'R'|$, as desired.

The following converse of Theorem G18 will also be useful; the proof is sufficiently straightforward to be left as an exercise.

Theorem G19 *If the alternate interior angles of a transversal with respect to a pair of distinct lines are equal, then the lines are parallel. The same is true of corresponding angles.*

To conclude this discussion of dilation, it would be pleasant to report that a composite of two dilations (with respective centers) is also a dilation (with perhaps some other center), but unfortunately such is not the case. An example will be given in an exercise below.

Exercises 6.2

1. (a) The dilation of a convex set is a convex set. (b) The dilation of a polygon is a polygon. (c) The dilation of a regular polygon is a regular polygon.

2. Prove Theorem G19.

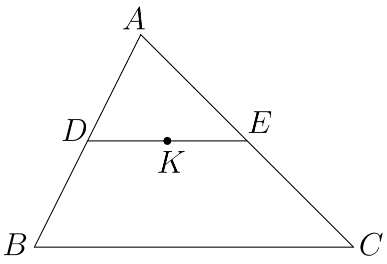
3. Let $ABCD$ and $A'B'C'D'$ be two quadrilaterals. Suppose there is a point K so that the rays R_{KA}, R_{KB}, R_{KC}, R_{KD} contain A', B', C', D', respectively. Assume also

$$\frac{|KA|}{|KA'|} = \frac{|KB|}{|KB'|} = \frac{|KC|}{|KC'|} = \frac{|KD|}{|KD'|}$$

If $ABCD$ is a square, then so is $A'B'C'D'$. (*Caution:* Be careful what you say and how you say it.)

4. Let O be a point not on a given circle $\mathcal{C}$ with center K. Let D be the dilation with center O and scale factor r. Prove that the image $D(\mathcal{C})$ is a circle, and that the center of $D(\mathcal{C})$ is the image under D of the center of $\mathcal{C}$. (*Caution*: This is a slippery proof. Follow the precise definitions of a circle and a dilation.)

5. Let D and E be the midpoints of AB and AC, respectively, of $\triangle ABC$, and let K be the midpoint of DE (see picture below). Let $\mathcal{D}$ be the dilation with center A and scale factor $\frac{1}{2}$. (a) If ϱ is the rotation of 180° around K, describe precisely the figure $\mathcal{D}(\varrho(\triangle ABC))$. (b) If T is the translation along $\overline{AD}$, describe precisely the figure $T(\mathcal{D}(\triangle ABC))$. (c) How are the figures in (a) and (b) related?



7. Let P and Q be two distinct points in the plane and let D_P, D_Q be two dilations with center at P, Q respectively, and with scale factor $\frac{1}{2}$ and 2, respectively. Prove that $D_P \circ D_Q$ is not a dilation. (*Hint*: Suppose $D_P \circ D_Q$ is equal to a dilation D_X with center X, then $D_P \circ D_Q$ maps X to X. On the other hand, there is no point Y so that $(D_P \circ D_Q)(Y) = Y$.)

8. Let P and Q be two distinct points in the plane and let D_P, D_Q be two dilations with center at P, Q respectively, and with scale factor r and s, respectively. If $rs \neq 1$, prove that there is a point X so that $(D_P \circ D_Q)(X) = X$. (In fact, in this case, the composition $D_P \circ D_Q$ is a dilation with center X, but the proof requires some tools that we have not yet developed.)

3 Similarity

Let $\mathcal{S}$ and $\mathcal{S}'$ be two sets in the plane.

Definition *We say $\mathcal{S}$ is* **similar** *to $\mathcal{S}'$, in symbols, $\boldsymbol{\mathcal{S} \sim \mathcal{S}'}$, if there is a dilation D so that*

$$D(\mathcal{S}) \cong \mathcal{S}'$$

More precisely, $\mathcal{S} \sim \mathcal{S}'$ means there is a congruence φ and a dilation D so that $\varphi \circ D$ maps $\mathcal{S}$ to $\mathcal{S}'$, i.e., $\varphi(D(\mathcal{S})) = \mathcal{S}'$. A composition $\varphi \circ D$ of a congruence φ and a dilation D is called a **similarity**. The **scale factor of the similarity** $\varphi \circ D$ is by definition the scale factor of the dilation D.

> *The fact that we define a similarity as a composite $\varphi \circ D$, where φ is a congruence and D is a similarity, is a matter of convention: we could have equally well defined a similarity by composing D and φ in the reverse order, i.e., $D \circ \varphi$. But of course, once so defined, one must be consistent throughout. One can prove that the two definitions are equivalent, in the sense that for any two sets $\mathcal{S}$ and $\mathcal{S}'$, $\varphi(D(\mathcal{S})) = \mathcal{S}'$ for some congruence φ and dilation D if and only if there is a congruence φ' so that $D(\varphi'(\mathcal{S})) = \mathcal{S}'$; see the Lemma at the end of this section.*
>
> *This situation is somewhat reminiscent of the definition of the multiplication of whole numbers, e.g., 3×5 can be defined either as $5+5+5$, or $3+3+3+3+3+3$, but once we fix the definition, we should not change it without explicitly invoking the commutativity of multiplication.*

We call attention to the fact that in the definition of similarity, a congruence (and not just an isometry) is used. Although every congruence is an isometry, *at this stage*, we still do not know whether an isometry is a congruence or not. So the advantage of a congruence over an isometry (at this stage) is that a congruence has an inverse and it also preserves the degrees of angles.

We will eventually prove that if $\mathcal{S} \sim \mathcal{S}'$, then also $\mathcal{S}' \sim \mathcal{S}$. Thus one can speak unambiguously of **two subsets being similar**. Since this proof is somewhat off the main line of reasoning, we will postpone this discussion until the end of this section.

Let us reflect a bit on the formal definition of similarity. It would not do to adopt a more restrictive definition of similarity by declaring simply that two figures are "similar" if one is the dilation of the other, because in our own minds, if after dilating a figure $\mathcal{A}$ to obtain figure $\mathcal{B}$ and $\mathcal{B}$ is congruent to figure $\mathcal{C}$, then we would still consider $\mathcal{A}$ and $\mathcal{C}$ to be "similar". Therefore we need the concept of congruence in order to give a legitimate definition of similarity. For this reason, *the concept of congruence must be firmly in place before we can discuss similarity.* At this point, you may wish to review the examples of dilation given in the preceding section, and reaffirm that there is a routine procedure to draw a figure similar to a given one (no matter what it is) with any specified scale factor.

We note explicitly that, although most of our attention will be lavished on triangles, this definition of similarity gives us a precise conception of what it means when we say one object (regardless of its shape) is similar to another. For example, it follows directly from the definition that all circles are similar to each other (see Exercise 5.3 below).

This concept of similarity applies not only to any geometric figure in the plane, but to figures in higher dimensions as well.

As in the case of congruence, the notation with the similarity of triangles, by tradition, is made to carry more information. We say $\boldsymbol{\triangle ABC \sim \triangle A'B'C'}$ if there is a similarity F so that

$$F(A) = A', \quad F(B) = B', \quad F(C) = C'$$

In other words, $\triangle ABC \sim \triangle A'B'C'$ means not only that there is a similarity F so that the sets $F(\triangle ABC)$ and $\triangle A'B'C'$ are equal, but that F specifically maps A to A', B to B', etc.

Theorem G20 *Given two triangles ABC and $A'B'C'$, their similarity, i.e., $\triangle ABC \sim \triangle A'B'C'$, is equivalent to the following equalities:*

$$|\angle A| = |\angle A'|, \quad |\angle B| = |\angle B'|, \quad |\angle C| = |\angle C'|$$

and

$$\frac{|AB|}{|A'B'|} = \frac{|AC|}{|A'C'|} = \frac{|BC|}{|B'C'|}$$

Remark It is in the proof of this theorem that we get to see why a similarity is defined as the composition $\varphi \circ D$ of a *congruence* φ (and not just an isometry) and a dilation D. It follows from Theorems G5 and G17 that a similarity preserves the degrees of angles, and this fact accounts for the validity of Theorem G20. We again emphasize that at this point we do not know if an isometry preserves the degrees of angles or not.

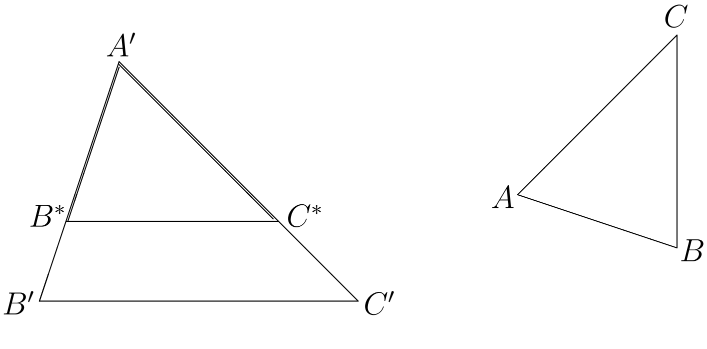
Proof If we have $\triangle ABC \sim \triangle A'B'C'$, then the assertions about angles and sides follow from Theorems G5 and G17. For the converse, we prove something stronger:

Theorem G21 (SAS for similarity) *Given two triangles ABC and $A'B'C'$. If $|\angle A| = |\angle A'|$, and*

$$\frac{|AB|}{|A'B'|} = \frac{|AC|}{|A'C'|}$$

then $\triangle ABC \sim \triangle A'B'C'$.

Proof of Theorem G21 If $|AB| = |A'B'|$, then the hypothesis would imply $|AC| = |A'C'|$ and we are reduced to the SAS criterion for congruence. Thus we may assume that $|AB|$ and $|A'B'|$ are not equal. Suppose $|AB| < |A'B'|$. Then the hypothesis that $|AB|/|A'B'| = |AC|/|A'C'|$ implies $|AC| < |A'C'|$. On $A'B'$, let B^* be the point so that $|A'B^*| = |AB|$. Similarly, on $A'C'$, let C^* be the point satisfying $|A'C^*| = |AC|$.



By SAS for congruence (Theorem G7), $\triangle A'B^*C^* \cong \triangle ABC$. Let φ be the congruence that maps $\triangle A'B^*C^*$ to $\triangle ABC$. Moreover, if r denotes the common value of $|AB|/|A'B'|$ and $|AC|/|A'C'|$, then the dilation D with center A' and scale factor r maps A' to A' of course, but also B' to B^* because by the definition of dilation, $D(B')$ is the point on the ray $R_{A'B'}$ so that the distance of $D(B')$ from the center A' is

$$r\,|A'B'| = \frac{|AB|}{|A'B'|}\,|A'B'| = |AB| = |A'B^*|$$

Thus $D(B') = B^*$. Similarly, $D(C') = C^*$. Thus D maps $\triangle A'B'C'$ to $\triangle A'B^*C^*$, thanks to Corollary 2 of Theorem G16. Hence, we have:

$$(\varphi \circ D)(\triangle A'B'C') = \varphi(D(\triangle A'B'C')) = \varphi(\triangle A'B^*C^*) = \triangle ABC$$

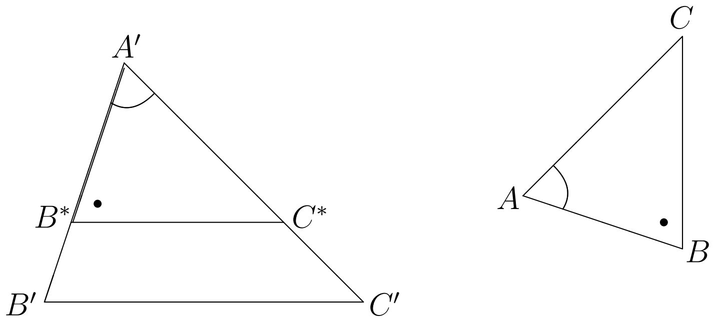
This shows that $\triangle A'B'C' \sim \triangle ABC$.

We next give the proof of the most easily applied criterion of similarity: AA for similarity.

Theorem G22 (AA for similarity) *Two triangles with two pairs of equal angles must be similar.*

REMARK Of course as soon as we prove that the sum of angles in a triangle is 180°, then knowing the equality of two pairs of angles is seen to be equivalent to knowing that all three pairs of angles are equal. This is why this criterion is sometimes cited as the *AAA criterion.*

Proof Let two triangles ABC and $A'B'C'$ be given. We may assume $|\angle A| = |\angle A'|$ and $|\angle B| = |\angle B'|$. Then we must prove that $\triangle ABC \sim \triangle A'B'C'$. If $|AB| = |A'B'|$, then the hypothesis would imply $\triangle ABC \cong \triangle A'B'C'$ because of the ASA criterion for congruence (Theorem G8). Thus we may assume that $|AB|$ and $|A'B'|$ are not equal. Suppose $|AB| < |A'B'|$. On $A'B'$, choose a point B^* so that $|A'B^*| = |AB|$, and let the line parallel to $B'C'$ and passing through B^* intersect $A'C'$ at C^*. By Theorem G18, $|\angle A'B^*C^*| = |\angle B'|$ which, by hypothesis, is equal to $|\angle B|$. Since also $|\angle A'| = |\angle A|$ by hypothesis, $\triangle A'B^*C^* \cong \triangle ABC$ by ASA for congruence. In particular, $|A'C^*| = |AC|$, by Theorem G5.



We now claim that

$$\frac{|A'B'|}{|A'B^*|} = \frac{|A'C'|}{|A'C^*|}$$

Once we prove this, we will see that the triangles $A'B'C'$ and ABC satisfy the conditions of SAS for similarity (Theorem G21) and are therefore similar. In order to prove the preceding equality, we take a point C_0 on $A'C'$ so that

$$\frac{|A'B'|}{|A'B^*|} = \frac{|A'C'|}{|A'C_0|}$$

By FTS, $L_{B^*C_0} \parallel L_{B'C'}$. Then the two lines $L_{B^*C_0}$ and $L_{B^*C^*}$ both have the property that they are parallel to $L_{B'C'}$ and they pass through B^*. By the Parallel Postulate, $L_{B^*C_0} = L_{B^*C^*}$, which implies $C_0 = C^*$. The equality that $|A'B'|/|A'B^*| = |A'C'|/|A'C_0|$ then becomes the sought-for equality $|A'B'|/|A'B^*| = |A'C'|/|A'C^*|$. The proof of Theorem G22 is complete.

Finally, we return to the discussion of $S \sim S'$ for two sets S and S'. By definition, this means that there is a dilation D and a congruence φ so that $\varphi(D(S)) = S'$. In this definition, the order of S and S' seems to matter, S first and S' second, because to say instead $S' \sim S$ would mean that there is a dilation D' and a congruence ψ so that $\psi(D'(S')) = S$. It is not obvious how to conclude from $\varphi(D(S)) = S'$ that $\psi(D'(S')) = S$ for some D' and ψ. Why is this so bad? Because if indeed it is the case that $S \sim S'$ and $S' \sim S$ do not hold simultaneously, then we cannot say "the two sets S and S' are similar", but must be careful to say "S is similar to S'" or "S' is similar to S" and make a distinction between the two. Life would then be unbearably complicated.

We salvage the situation by proving that $S \sim S'$ must imply $S' \sim S$. In standard terminology, this says **similarity is a symmetric relation**. The crux of the matter is the following lemma:

Lemma *Let D be a dilation and φ a congruence. Then there is a dilation D' so that $D \circ \varphi = \varphi \circ D'$.*

This is *almost* the statement that D and φ commute, but it doesn't quite say that because D' is not going to be D in general. In fact, it is not difficult to see what D' must be if the lemma is true. Indeed, $D \circ \varphi = \varphi \circ D'$ implies that $D' = \varphi^{-1} \circ D \circ \varphi$. This then tells us how to define D' in order to prove the Lemma. All it remains is therefore to prove that this D' so defined must be a dilation. Here is a hint: If D has center O and scale factor r, let $\varphi(O') = O$. Then show that D' has center O' and scale factor r. We leave the details of the proof of the Lemma and also the proof that $S \sim S'$ must imply $S' \sim S$ to an exercise.

Exercises 6.3

1. Let D, E, F be the midpoints of the sides BC, AC, AB, respectively, of a triangle ABC. Then $\triangle DEF \sim \triangle ABC$ with a scale factor of 2.

2. [*This problem generalizes problem 6 of Exercises 5.1.*] *Assume FTS.* Let L_1, L_2, and L_3 be three mutually parallel lines, and let ℓ and ℓ' be two distinct transversals which intersect the three parallel lines at A_1, A_2, A_3, and B_1, B_2, B_3, respectively. Then

$$\frac{|A_1A_2|}{|A_2A_3|} = \frac{|B_1B_2|}{|B_2B_3|}$$

3. Prove that all circles are similar to each other. (*Caution:* Don't skip steps.)

4. Prove that two rectangles are similar to each other if and only if either the ratios of their sides are equal or the product of these ratios is 1. Precisely, let the lengths of the sides of one rectangle be a and b, and those of the other be a' and b'; then the rectangles are similar if and only if either $\frac{a}{b} = \frac{a'}{b'}$ or $\frac{a}{b} \cdot \frac{a'}{b'} = 1$.

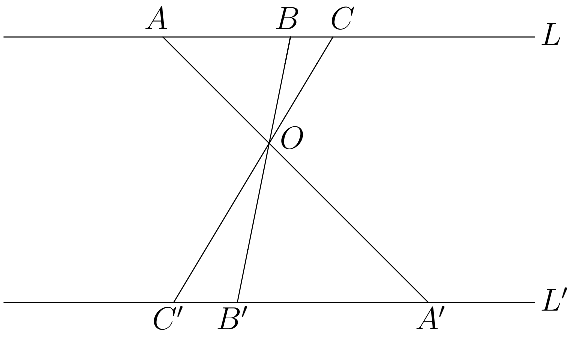
5. Let L and L' be two lines intersecting at a point O. Take any point P on L, and let the line passing through P and perpendicular to L' meet L' at a point P'. Then the ratio $\frac{|PP'|}{|OP'|}$ is independent of the position of P on L, i.e., if Q is another point on L, and if the line passing through Q and perpendicular to L' meets L' at a point Q', then

$$\frac{|PP'|}{|OP'|} = \frac{|QQ'|}{|OQ'|}$$

6. (a) **(AAS)** Suppose two triangles have two pair of equal angles. If they have a pair of equal sides, then they are congruent. (b) Let $|\angle B| = |\angle C|$ in $\triangle ABC$. Then $|AB| = |AC|$. (c) Every point on the angle bisector of an angle is equi-distant from both sides of the angle.

> (*Note:* In some sense, these three assertions should be proved in the setting of congruence, not *similarity*; any theorem related to similar triangles requires the FTS, which is a more sophisticated theorem than anything about congruent triangles. That said, the virtue of this problem is that you get to see at least another approach to these standard facts.)

7. Suppose we have two parallel lines L and L', and a point O not lying on either line. Let three lines passing through O intersect L and L' at points A, B, C, and A', B', C', respectively.



(This picture puts O between L and L', but O could be anywhere.) Prove that

$$\frac{|AB|}{|A'B'|} = \frac{|BC|}{|B'C'|} = \frac{|AC|}{|A'C'|}$$

8. Suppose you are a teacher in middle school and you are handed a textbook series that takes up similarity in grade 7 and congruence in grade 8. (Such a series exists.) (a) Do you believe such a curricular decision is defensible? Explain. (b) If you are a seventh grade teacher, what would you do? (Obviously there will be no unique answer to part (b), but the idea is that you had better start thinking about such real-world situations because the world out there is full of curricula that make no sense, and your ability to adjust is, alas, part of your responsibility.).

9. (a) Write out a detailed proof of the Lemma at the end of this section. (b) Write out a complete proof of the fact that, for any two sets S and S', if $S \sim S'$, then $S' \sim S$.

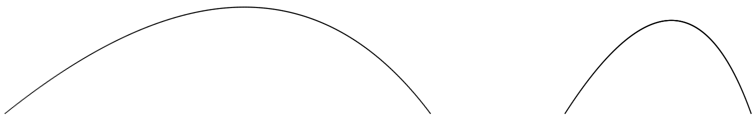
10. Suppose for three sets S_1, S_2, and S_3, we have $S_1 \sim S_2$, and $S_2 \sim S_3$. Prove that $S_1 \sim S_3$. (In standard terminology, this says **similarity is a transitive relation.**)

Chapter 7: Length and Area

1 The concept of measurement

In this section, we want to give a general overview of the subject of **geometric measurements**: length, area, and volume. Very roughly, a geometric measurement assigns a number to a geometric figure that serves to indicate the "size" of the figure relative to that particular measurement. For example, a general curve will have a positive length and a general planar region will have positive area, but the same curve will have zero area and the same region will have zero volume. Each geometric measurement is thus an **assignment** of a number to a geometric figure. Can this assignment be done at will? If not, what are the guiding principles that help decide what it is? These are the questions we want to discuss in this section.

Length, area, and volume come up naturally in normal conversations and are routinely used in all phases of daily life. For this reason, the corresponding mathematical definitions carry an additional burden: they must prove their worth by producing measurements in familiar situations that are consistent with this common knowledge. Take the case of length, for instance. To each curve C, we would like to assign a number $|C|$ so that, if C is one of the common curves such as a square or a circle, then $|C|$ *is* the length of C as we know it. Thus, if we denote the length of a curve C by $|C|$, this number $|C|$ must be consistent with our intuition of what "length" should be. Let us amplify on the last statement: the length function clearly cannot be randomly defined because people would not take kindly to a function that assigns to the following curve on the left a "length" that is smaller than that for the curve on the right even if they cannot articulate, precisely, what "length" means:



Therefore we are going to formulate set of characteristic properties that such an assignment is expected to possess, and then we start from scratch — very much in the spirit of §1 of Chapter 5 — by looking at the simplest curves. We will determine what length each one should have in view of these characteristic properties, and then go on to more complicated curves and decide anew what lengths these should be assigned.

And so on. We will do the same with area.

Mensuration formulas for length, area and volume belong to the oldest part of mathematics, and for a good reason. They met some basic human needs at the dawn of civilization, such as measuring land for farming and measuring grains for bartering. The earliest mathematical records of the oldest civilizations — Babylonian, Egyptian, Chinese, and Indian — contain area formulas of rectangles and triangles. These formulas are therefore part of the staple of the school mathematics curriculum. Unfortunately, they also belong to the most misunderstood part of the school curriculum. One reason is that although the concepts of length, area, and volume seem to be straightforward in an intuitive sense, they become quite complex with any attempt at a more precise understanding. Students need careful explanations of these concepts in ways that are grade-level appropriate as well as mathematically sensible. Because most school curricula rarely rise to the challenge, students are left with the proverbial concept of it area, for example, as *length times width*. On a more sophisticated level, most students believe that they know what the number π is because it is just circumference divided by diameter. They do not stop to reflect that they have no way of explaining what "circumference" is other than to put a string around a cylinder.

This chapters will try to elucidate the concepts of length and area. We will try to navigate a middle course between what is correct and what is pedagogically feasible for middle school students. In general, we will restrict the discussion to the most common geometric objects, which are fortunately very well-behaved in the sense that the curves are never too wriggly and the surfaces are never too rugged, so that we can stay on the intuitive level without having to attend to some serious logical difficulties. You are therefore forewarned that some technical details will be missing in this chapter, but that the main ideas are nevertheless correct.

The first, and critical step in performing any measurement is the choice of a **unit**, which in the context of geometry is, by definition, the geometric object to which we assign the numerical value 1.[63] Once such a unit has been chosen, the length or area

[63] This echoes the corresponding situation in the definition of the number line: we have to fix a choice of the point 1 (in addition to 0) before we can fix the positions of the remaining whole numbers.

of every geometric figure becomes a "comparison" of this figure with the unit. How to do this "comparison" is then our main concern.

Limiting ourselves now to the specific measurements of length, area, and volume, we state some general principles governing these measurements. These principles are supposed to be so self-evident that you all subscribe to them. (Compare the discussion of (L1)–(L6) in §1 of Chapter 5.) There are four of them, and we call them the **Fundamental Principles of Geometric Measurements.** We start with a fixed **collection** of geometric figures, be they curves, planar regions, or solids in space. To each figure G in this collection, we assign a number $|G|$, called its **geometric measurement**; if G is a curve, $|G|$ is its length, if G is a planar region, $|G|$ is its area, and if G is a solid, $|G|$ is its volume.

(M1) There is a fixed figure G_0 in the collection, to be called the **unit figure**, so that $|G_0| = 1$. In more detail:

> For length, the unit figure is the unit segment, i.e., $[0, 1]$.
>
> For area, the unit figure is the unit square, i.e., a square whose sides are of length 1.
>
> For volume, the unit figure is the **unit cube**, i.e., a rectangular solid all of whose sides have length 1.

(M2) If a figure A in the collection has geometric measurement a, and a figure B in the collection is congruent to A, then the geometric measurement of B is also a. In other words, length, area or volume is the same for congruent figures.

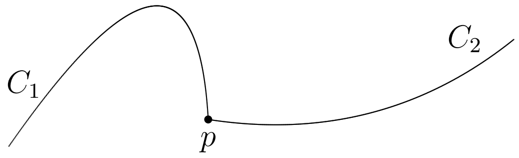
In view of (M2), we will adopt the usual abuse of language and also *call any segment, rectangle, and rectangular solid that is congruent to the unit segment, unit square, and unit cube, respectively, a unit segment, unit square, and unit cube, respectively.*

It is a good idea in teaching to bring out the direct relevance of congruence to the discussion of length, area, and volume. A common problem in school mathematics is to make students learn some concepts but fail to show what these concepts are good for. In this case, students get to see that *congruence* is more than a fancy way to say "same size and same shape". Rather, it deserves to be learned because *it lies at the foundation of something that matters to them: geometric measurements.*

(M3) (Additivity) Geometric measurement is **additive** in the sense that if a figure G is the union of two other figures in the collection, G_1 and G_2, so that the intersection $G_1 \cap G_2$ is contained in the boundaries of G_1 and G_2, then the geometric measurement of G is the sum $|G| = |G_1| + |G_2|$. More precisely:

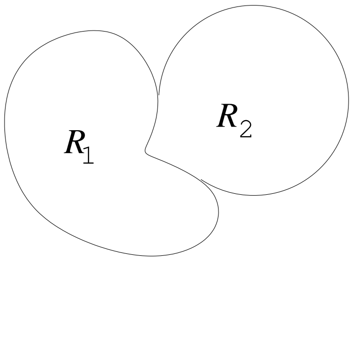
If two curves intersect at at only their endpoints and their lengths are known, then the length of their union is equal to the sum of their lengths.

> Thus the length of the curve below obtained by joining the curve C_1 and the curve C_2 at the point p is the sum of the length of C_1 and the length of C_2:



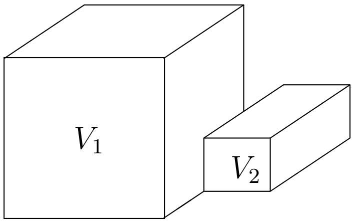
If two planar regions intersect only at (part of) their respective boundary curves, then the area of their union is equal to the sum of their areas.

> Thus the area of the region below, which is the union of R_1 and R_2, is equal to the sum of the area of R_1 and the area of R_2:

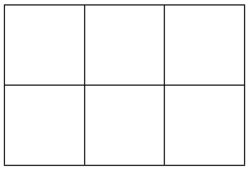


If two solids in 3-space intersect only at (part of) their respective boundary surfaces, then the volume of their union is equal to the sum of their volumes.

> Thus, for example, the volume of the solid which is the union of the two rectangular solids V_1 and V_2, with parts of their boundaries in common, is the sum of the volume of V_1 and the volume of V_2:



We should not fail to point out that (M3) is what lies behind how we teach area in elementary school. For example, we have a rectangle whose sides have lengths 2 and 3. We can therefore draw the following picture:



Ideally, a teacher in 4th grade would point out that the area of this rectangle, *according to* (M3), is the sum of the areas of the 2 rows of 3 unit squares; because of what students have just learned about whole-number multiplication, there are 2×3 such unit squares. Therefore, the area of the rectangle is equal to 2×3.

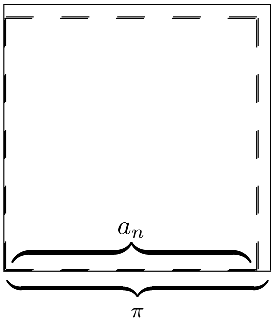
There is a fourth property that is equally basic, but which is more sophisticated and, at the same time, more difficult to articulate precisely. We are going to announce it in the following tentative form, with the understanding that it will be further clarified in each discussion of length and area.

(M4) Given a geometric figure G in the collection. Suppose $\{G_n\}$ is a sequence of geometric figures in the same collection such that $\{G_n\}$ *converges* to G, in a sense to be made precise. Then $|G_n| \to |G|$.

The meaning of "convergence" in the sense of numbers or geometric figures will be taken in the intuitive sense, and will usually be very transparent in context. The naive content of (M4) is so appealing that we can give a simple illustration, *using informal language*, of the basic idea involved in the case of area. Suppose we have a square S whose side has length π and we want to know the area $|S|$ of S. Now if the length of the side is a fraction, say $\frac{22}{7}$ instead of π, then what we learned from Theorem 1 and 2 in §4 of Chapter 1 is that the area of S must be $\frac{22 \times 22}{7 \times 7}$, which is approximately 9.87755, and we are done. But π is not a fraction, so have to rely on the validity of (M4) to compute the area of this square. We get an increasing sequence of *fractions* (a_n) so that $\lim a_n = \pi$. For example, since there is a decimal expansion of π,

$$\pi \;=\; 3.14159\;26535\;89793\;23846\;26433\;83279\ldots$$

we may let $a_1 = 3.1$, $a_2 = 3.14$, $a_3 = 3.141$, $\ldots a_{14} = 3.14159\;26535\;8979$, and in general, $a_n =$ the first (from the left) $n+1$ digits of the decimal expansion of π. In any case, the explicit value of each member of the sequence is immaterial and what is important is that we have a sequence of increasing *fractions* converging to π. Then let S_n be the square whose side has length a_n. We may picture S_n as the dotted square in the following:



As $n \to \infty$, the boundary of S_n gets arbitrarily close to the boundary of S because $a_n \to \pi$, and S_n fills up S, so that it would be reasonable to describe this phenomenon as "S_n *converges* to S". Intuitively, the "area of S" *is* the limit of the areas $|S_n|$. Since the area of S_n is $(a_n)^2$, we can easily believe that

$$\lim_n |S_n| = \lim_n (a_n)^2 = \lim_n a_n \; \lim_n a_n = \pi \cdot \pi = \pi^2$$

Therefore the area $|S|$ of S is $\pi \cdot \pi = \pi^2$, and the main substance of (M4) in this special case is to guarantee that this intuitive understanding is correct. Needless to say, π^2 *is* what we normally consider to be the area of S.

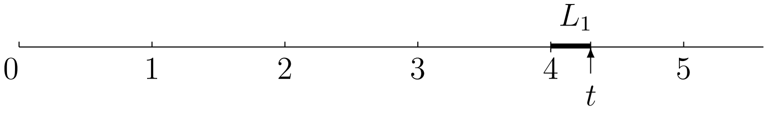
2 Length

Length of a segment

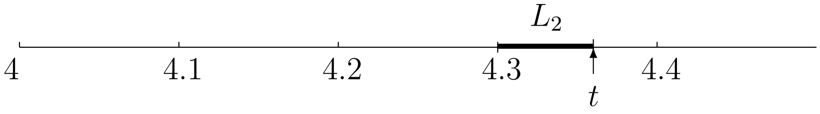
We begin with the measurement of the lengths of the simplest curves: line segments. We know that the length of the unit segment is 1, by (M1). By (M2), the length of any segment congruent to $[0, 1]$ is also 1. Now let AB be an arbitrary segment in the plane[64]. How to determine the length $|AB|$ explicitly? It is necessary to bring out the fact that we are not interested in knowing the measurement *in principle*. Were that the case, use a congruence to bring AB to the number line so that A is at 0, then the point B falls on some number t and of course we know from (M2) that $|AB| = t$. But t is not the answer to the problem until we can prescribe an algorithm to determine what t is, precisely. In the subject of geometric measurements, the goal is always to have explicit determination of the measurement. This is the reason for the various length, area, and volume formulas in the subject.

Back to AB, let us say t falls between 4 and 5. Let the segment $[4, t]$ be denoted by L_1:

[64] Or in 3-space if we expand our horizon to 3-space. There is no difference in the reasoning.



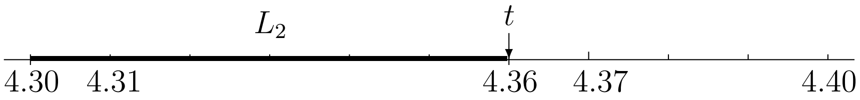
By the principle of additivity (M3), we have $t = 4 + |L_1|$. We have to measure the length of L_1. Because the unit segment is too long, we have to introduce a smaller unit for this purpose. By common convention, the new unit to use in this context is 0.1, or what is the same thing, the length one part in a partition of the unit segment into 10 congruent parts. Thus we have the division points 4.1, 4.2, ..., 4.9. Let us say t falls between 4.3 and 4.4. Denote the segment $[4.3, t]$ by L_2. After magnification, we get a picture that looks like this:



Since L_1 is $[4, t]$, by (M3) again, we have $|L_1| = 0.3 + |L_2|$, and therefore

$$|AB| \; = \; 4 + 0.3 + |L_2|$$

Since $|L_2|$ is smaller than 0.1, we now handle L_2 in exactly the same way: we measure it by using a yet smaller unit, which is 0.01, i.e., the length of one part when $[4.3, 4.4]$ is partitioned into 10 congruent parts. Suppose t falls *exactly* on the 6th division point after 4.3. After magnification again, the picture is the following:

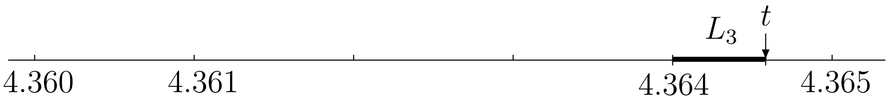


This means geometrically that exactly 6 of the 0.01-units can be fitted into L_2, so that by (M3), $|L_2| = 6 \times 0.01 = 0.06$. We therefore have

$$|AB| \; = \; 4 + 0.3 + 0.06,$$

which is of course equal to 4.36.

Of course, the measurement process may never end, in the sense that t may never fall exactly on any of the division points corresponding to units of lengths 0.01, 0.001, 0.0001, . . . For example, in the preceding example, t could have fallen between 4.364 and 4.365, with a segment L_3 left over:



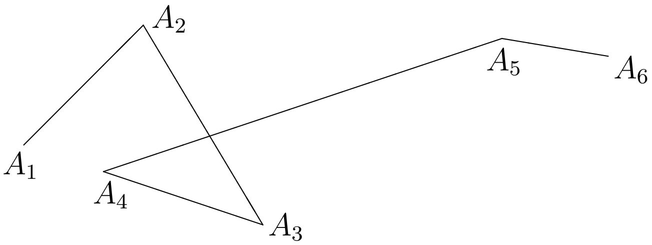
By (M3) again,

$$|AB| \;=\; 4 + 0.3 + 0.06 + +|L_3|,$$

with $|L_3| < 0.001$, and the process continues. If it never ends, then the preceding description yields an algorithm that gives the length of AB as an infinite decimal.

Lengths of polygonal segments

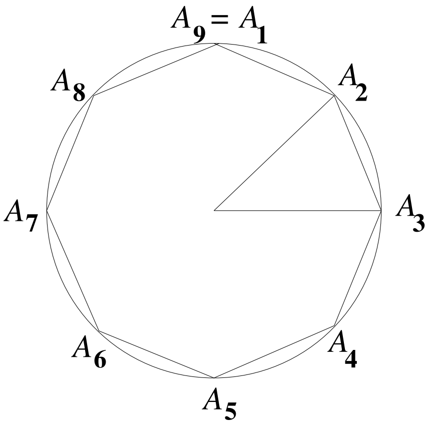
Having described how to measure segments, we proceed to measure something on the next level of complexity. By a **polygonal segment** $A_1A_2 \cdots A_n$, we mean a sequence of segments A_1A_2, A_2A_3, . . . $A_{n-1}A_n$, with the understanding that these segments need not be collinear and intersections among them are allowed. The points A_1, A_2, . . . A_n are called the **corners** of $A_1A_2 \cdots A_n$. We will limit our discussion to polygonal segments in the plane, but everything we say in fact makes perfect sense in 3-space.



In accordance with (M3), the **length** $|A_1A_2 \cdots A_n|$ of $A_1A_2 \cdots A_n$ should be defined as the sum

$$|A_1A_2 \cdots A_n| = |A_1A_2| + |A_2A_3| + \cdots + |A_{n-1}A_n|.$$

In the case of an n-gon, it is the polygonal line $A_1A_2 \cdots A_nA_{n+1}$, where $A_{n+1} = A_1$ (but with the additional condition that the sides do not intersect each other except at the endpoints between consecutive sides). The length of this polygonal segment is just the sum of the lengths of all the sides of the polygon, i.e., the perimeter of the polygon. Here is a regular 8-gon **inscribed in a circle**, in the sense that all the vertices of the polygon lie on the given circle.



In any case, we know how to compute the lengths of all polygonal segments at this point.

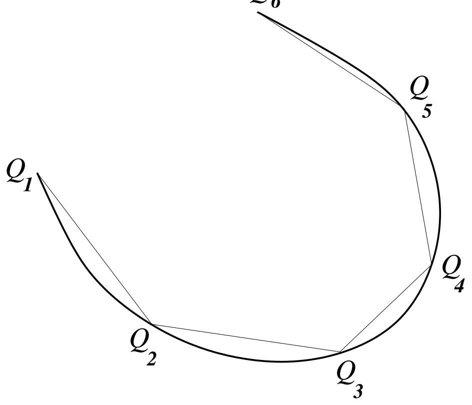
Lengths of curves

Polygonal lines, with the exception of a finite number of corners, are linear (i.e., straight) objects. We now must confront a general, non-straight curve. We have not given a precise definition of what a "curve" is, and will not attempt to do so because, incredible as it may seem, it is a complicated mathematical concept. In school mathematics, the curve of greatest interest is without doubt the circle. If you keep the circle in mind whenever we talk about a curve, and you will not be too far wrong. We shall therefore proceed on an intuitive level where curves are concerned, but everything to be said will be correct as soon as some technical details are in place.

Most curves are not polygonal lines, e.g., a circle or an ellipse. To determine the length of curves in general, we are guided by (M4) to adopt the following technique of measurement:

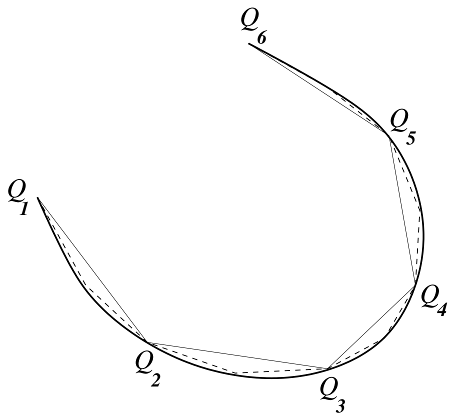
> Extend our knowledge from the known (polygonal lines) to the unknown (general curves) by approximating the unknown quantity with a sequence of the known quantities.

In greater detail, this means the following. Since we already know how to compute the lengths of polygonal segments, we will use these to compute the length of any piecewise smooth curve. The basic idea is that a curve C can be approximated by polygonal segments whose vertices lie on C. We say a polygonal segment $P = Q_1Q_2 \cdots Q_n$ is a **polygonal segment on $\boldsymbol{C}$** if its ordered sequence of vertices Q_1, Q_2, …, Q_n belong to C and Q_1 and Q_n are the endpoints of C. Here is an example with $n = 6$:



Now the basic (and trivial) observation is that the approximation of C by such a polygonal segment would improve if the distance between each pair of adjacent vertices decreases. We illustrate this fact by drawing a new polygonal segment on C (the dotted one) with only a single vertex inserted between any two of the preceding Q_i's.

It is now intuitively clear that if a polygonal segment on a given C has the property that the distance between any pair of adjacent vertices of the polygonal segment is extremely small, then the polygonal segment would become almost indistinguishable from C itself. Therefore to get a good approximation of a given curve by use of polygonal segments on it, we have to make sure that the distance between *every* pair of adjacent vertices is small. One way to do this is is to specify that the **mesh** of a polygonal segment $P = Q_1Q_2 \cdots Q_n$ (in symbol: $\boldsymbol{m(P)}$) is small, where, by definition, $m(P)$ is the *maximum* of the lengths $\{|Q_1Q_2|, |Q_2Q_3|, \ldots, |Q_{n-1}Q_n|\}$.



Thus if $m(P)$ is small, then the distance between any pair of adjacent vertices of P, being at most equal to $m(P)$, must be small as well.

Let $\{P_n\}$ be a sequence of polygonal segments on a curve C. We say $\{P_n\}$ **converges to $\boldsymbol{C}$** (in symbol: $\boldsymbol{P_n \to C}$) if $m(P_n) \to 0$. A few more experiments with such polygonal segments would convince you that it is entirely reasonable to **define** the length $|C|$ of a curve C to be the "limit" of $|P_n|$ as n gets increasingly large, if $P_n \to C$. In symbols:

$$|C| \overset{\text{def}}{=} \lim_{n\to\infty} |P_n| \qquad \text{if } \; P_n \to C. \tag{9}$$

There is a precise technical meaning of "limit", but here it suffices to understand it in the intuitive sense. But note that in a real sense, you are already used to seeing the concept of limit in action: the above description of how to measure the length of a segment by successively measuring the left-over segments L_1, L_2, L_3, ...to arrive at an infinite decimal is in fact a limiting process. In this case, the value of $|AB|$ *is* the limit of the sequence of finite decimals 4, 4.3, 4.36, 4.365, ...Moreover, it is a theorem (that we shall not be able to prove here) that this limit does not depend on the choice of the sequence of P_1, P_2, ... *so long as the maximum distances between adjacent corners of P_n get arbitrarily small as n gets large.* This freedom in the choice of these P_1, P_2, ...is important: it means that in a given geometric situation, we can judiciously pick a sequence of points P_1, P_2, ...to facilitate the evaluation of the limit in equation (9). We will see an example of this in the next sub-section.

We have elected to omit a technical detail from the definition (9) of curve length, namely, that the class of curves under consideration is what is known as the **rectifiable curves**, which are exactly the curves for which the limit in definition (9) always makes sense regardless of the choice of the approximating polygonal lines. The curves you encounter in daily life are almost certainly rectifiable.

Circumference of a circle

We illustrate the discussion in the preceding sub-section by considering the length of a circle of radius r, which is usually called the **circumference** of the circle. Let such a circle be denoted by $\mathbf{C}(r)$. Among all possible approximating sequences of polygonal lines to the circle, one is distinguished, namely, the sequence of regular n-gons inscribed in this circle as n gets large.[65] Let R_n denote the perimeter of a regular n-gon inscribed in this circle. Then the definition in (9) when specialized to this situation becomes:

$$|\mathbf{C}(r)| \;=\; \lim_{n\to\infty} R_n$$

We digress to consider the effect of a dilation on length. Let $\mathcal{D}_s$ be a dilation of scale factor s around a point O. Then every segment L is dilated to a segment $\mathcal{D}_s(L)$. According to Theorem G17 of Chapter 6, $|\mathcal{D}_s(L)| = s|L|$. If now P is a polygonal segments $A_1A_2\cdots A_n$, then $\mathcal{D}_s(P)$ is the sequence of segments $\mathcal{D}_s(A_1A_2)$, $\mathcal{D}_s(A_2A_3)$, $\ldots$, $\mathcal{D}_s(A_{n-1}A_n)$, so that,

$$\begin{aligned}
|\mathcal{D}_s(P)| &= |\mathcal{D}_s(A_1A_2)| + |\mathcal{D}_s(A_2A_3)| + \cdots + |\mathcal{D}_s(A_{n-1}A_n)| \\
&= s|A_1A_2| + s|A_2A_3| + \cdots + s|A_{n-1}A_n| \\
&= s(|A_1A_2| + |A_2A_3| + \cdots + |A_{n-1}A_n|) \\
&= s|P|.
\end{aligned}$$

In other words, for any polygonal line P,

$$|\mathcal{D}_s(P)| = s\,|P| \tag{10}$$

If C is any curve and P_n is an approximating sequence of polygonal lines of the curve $\mathcal{D}_s(C)$, then by the definition (9), $|\mathcal{D}_s(C)| = \lim_{n\to\infty} |\mathcal{D}_s(P_n)|$. Taking (10) into account, we have

$$|\mathcal{D}_s(C)| = s \lim_{n\to\infty} |P_n| = s|C|\,,$$

[65] The proof that the sequence of inscribed regular n-gons approximates the circle in the sense that its mesh decreases to 0 is tedious. We will not spend time on such a proof at this juncture.

where the last step uses the definition in (9) again. Thus for any curve C,

$$|\mathcal{D}_s(C)| = s\,|C| \tag{11}$$

Now if C is $\mathbf{C}(r)$, a circle of radius r, then you can convince yourself that $\mathcal{D}_r(\mathbf{C}(1)) = \mathbf{C}(r)$ for any r (see problem 4 in Exercises 6.2). Therefore, by (11),

$$|\mathbf{C}(r)| = r|\mathbf{C}(1)| \tag{12}$$

Recall that $\mathbf{C}(1)$ is called the **unit circle**. We see that to measure the circumference of any circle, it suffices to measure the circumference of the unit circle. At this point, it would be legitimate to *define* the number π to be $\frac{1}{2}|\mathbf{C}(1)|$, but because the limit in (9) does not lend itself very well to experimentation, we prefer to define π differently as the area of the unit circle. This will be done in the next section.

Up to this point, the advantage of choosing the perimeter of an inscribed regular n-gon on the circle as an approximating sequence of polygonal segments on the circle is not apparent, but it will be very apparent when we simultaneously deal with circumference and the area of (the region enclosed by) the circle in the next section.

3 Area

The concept of area

Standard area formulas

Dilation and area

Area of general regions

π and the area of a circle

Geometric approximations of π

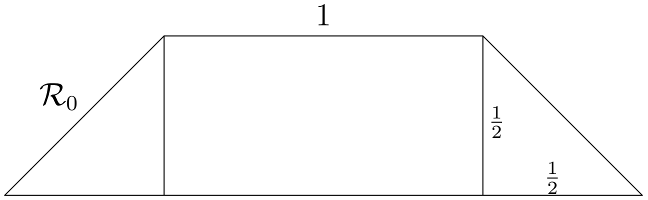
The measurement of area is special among geometric measurements in dimensions greater than 1 in that the area of the most common rectilinear figures (i.e., polygons) can be computed *exactly* without the use of limits provided we assume that the area of a rectangle is the product of the (lengths of the) sides. The area formulas of triangles, trapezoids, etc., lie at the core of the school mathematics curriculum. Accordingly, they occupy a place of honor in this section. A second noteworthy feature of this section is the clarification of the number π and the relationship between the circumference of a circle and the area of (the region enclosed by) a circle.

The discussion of area in this section takes place in a given plane.

Concept of area

There is a wide latitude in the choice of a unit for the measurement of area, but the standard unit figure is the *unit square*, i.e., the square whose sides are all of length 1. Nothing else has comparable simplicity and ease of having several of them "packed" together to fill a region. If the unit of length is an inch or a cm (centimeter), then it is customary to call the area of the unit square a **square inch** or **square cm**, respectively. The reason for the nomenclature is self-explanatory. In the following discussion, however, we will omit any reference to the *name* of the explicit unit of length being used (inch or cm or whatever), and will in particular not mention the unit of area as square inch or square cm, or in fact any other name.

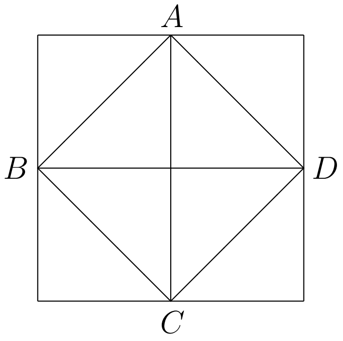
By (M3), the area of a given region $\mathcal{R}$ in a plane Π (which will be fixed from now on) is, intuitively, the number of unit squares, or fractional parts of a unit square, that can **tile** $\mathcal{R}$ in the sense that $\mathcal{R}$ can be expressed as the union of a number of these squares or fractional parts of squares which intersect each other at most at (part of) the edges. As illustration, we saw in §1 how (M3) leads to the computation of the areas of rectangles whose sides have whole-number lengths. As another illustration, let us compute the area of the following region $\mathcal{R}_0$ consisting of a rectangle in the middle with width 1 and height $\frac{1}{2}$, and two congruent isosceles right triangles with each **leg** (a side not facing the right angle) having length $\frac{1}{2}$. Please remember that at this point, *we do not know any area formulas for triangles or rectangles* so that this computation has to be carried out using only (M1) – (M3).



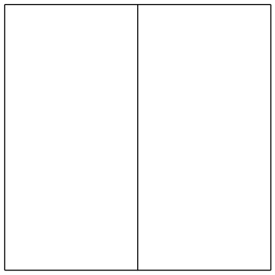
By (M3),

area of $\mathcal{R}_0$ = area of two triangles + area of middle rectangle

We now use (M2) to find out the individual areas on the righthand side. Because the unit square has area 1, the following figure shows a partition of the unit square into 8 congruent right triangles with A, B, C, D being the midpoints of the respective sides (for the asserted congruence, see problem 2 in Exercises 5.4).



By assumption (M2), all 8 triangles have equal area. Thus the partition divides the unit square into 8 equal parts (in terms of area), so that by the definition of a fraction, each triangle represents $\frac{1}{8}$ (of unit area). The two triangles of $\mathcal{R}_0$, each being congruent to each of these 8 triangles, therefore have area $\frac{1}{8}$ each (by (M2) again). Moreover, the following simple division of the unit square into two parts by joining the midpoints of two opposite sides



gives rise to 2 congruent rectangles, each congruent to the rectangle in $\mathcal{R}_0$. Again by (M2), the two rectangles have equal area and therefore each is $\frac{1}{2}$ in terms of the unit (square). By (M2), this means the rectangle in $\mathcal{R}_0$ has area $\frac{1}{2}$. Putting all these together, we have:

$$\begin{aligned} \text{area}(\mathcal{R}_0) &= \text{area(two triangles)} + \text{area(rectangle)} \\ &= (\frac{1}{8} + \frac{1}{8}) + \frac{1}{2} \;=\; \frac{3}{4} \end{aligned}$$

Standard area formulas

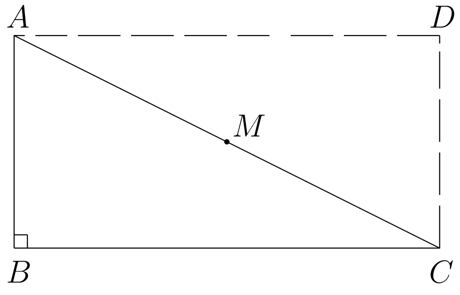
The whole discussion in this sub-section hinges on the simple statement that

$$\textit{area of rectangle} \;=\; \textit{product of the (lengths of the) sides}$$

The validity of this formula when both sides have fractional lengths follows from Theorem 2 in §4 of Chapter 1. By FASM, this formula remains valid even when the sides of the rectangle have arbitrary lengths.

It is astonishing how much useful information can be extracted from this simple formula alone. We will show how to exploit this area formula to compute the areas of triangles, parallelograms, trapezoids, and in fact any polygon at least in principle.

We begin with triangles. Consider a right triangle $\triangle ABC$ with $AB \perp BC$. We compute its area by expanding it to a rectangle. Let M be the midpoint of AC.



We observe that if ρ denotes the rotation of 180° around M so that $\rho(C) = A$, $\rho(A) = C$ and $\rho(B) = B$, then the quadrilateral $ABCD$ is in fact a rectangle, for the following reason. From $\rho(BC) = AD$ we conclude $AD \parallel BC$, and from $\rho(AB) = CD$ we conclude $AB \parallel CD$ (Theorem G1, Chapter 5). This shows $ABCD$ is a parallelogram. Since $AB \perp CD$, we also have $AB \perp AD$ because a line perpendicular to one of two parallel lines is perpendicular to both (Theorem G3, Chapter 5) . So $|\angle BAD| = 90°$. Similarly, $|\angle ADC| = |\angle DCB| = 90°$, and $ABCD$ is a rectangle. Now, the rotation ρ maps $\triangle ABC$ to $\triangle CAD$, (M2) implies that the two triangles have the same area. Thus the usual argument with (M3) proves that

$$\text{area}(\triangle ABC) \;=\; \frac{1}{2}\,\text{area}(ABCD).$$

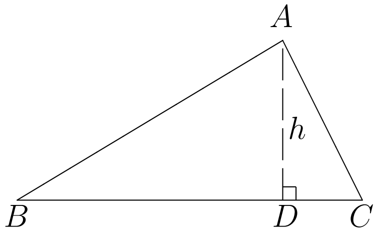
By the area formula of a rectangle, we get,

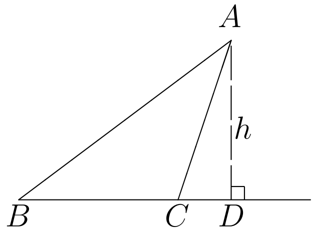
$$\text{area}(\triangle ABC) \;=\; \frac{1}{2}\,|AB| \cdot |BC|.$$

Recall that AB and BC are called the legs of $\triangle ABC$. We therefore have:

$$\textit{area of right triangle} \;=\; \frac{1}{2} \cdot \textit{product of (the lengths of) its legs}$$

Next, let $\triangle ABC$ be arbitrary and let AD be the **altitude** from the vertex A to BC (i.e., the segment which joins A to the line BC and is perpendicular to line L_{BC}). We now obtain two *right* triangles, $\triangle ABD$ and $\triangle ACD$, so that the preceding formula becomes applicable. Then there are two cases to consider: D is inside the segment BC, and D is outside BC. See the figures:





In either case, AD is called the **height** with respect to the **base** BC. By the usual abuse of language, **height** and **base** are also used to signify the *lengths* of AD and BC, respectively. With this understood, we shall prove in general that

$$\textit{area of triangle} \;=\; \frac{1}{2}\,(\textit{base} \times \textit{height})$$

For convenience, we shall use h to denote $|AD|$. Then this is the same as proving

$$\text{area}(\triangle ABC) \;=\; \frac{1}{2}\,|BC|\cdot h$$

In case D is inside BC, we use (M3) and refer to the figure above to derive:

$$\begin{aligned}\text{area}(\triangle ABC) &= \text{area}(\triangle ABD) + \text{area}(\triangle ADC)\\ &= \frac{1}{2}\,|BD|\cdot h + \frac{1}{2}\,|DC|\cdot h\\ &= \frac{1}{2}\,(|BD| + |DC|)\;h\\ &= \frac{1}{2}\,|BC|\cdot h\end{aligned}$$

In case D is outside BC, we again use (M3) and refer to the figure above to obtain:

$$\text{area}(\triangle ABD) \;=\; \text{area}(\triangle ACD) + \text{area}(\triangle ABC)$$

This is the same as

$$\frac{1}{2}\,|BD|\cdot h \;=\; \frac{1}{2}\,|CD|\cdot h + \text{area}(\triangle ABC).$$

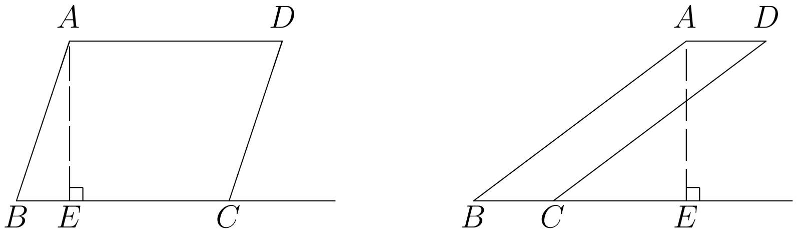
Therefore,

$$\begin{aligned}\text{area}(\triangle ABC) &= \frac{1}{2}\,|BD|\cdot h - \frac{1}{2}\,|CD|\cdot h\\ &= \frac{1}{2}\,(|BD| - |CD|)\;h\\ &= \frac{1}{2}\,|BC|\cdot h\end{aligned}$$

Thus the area formula for triangles has been completely proved.

Most textbooks mention the first case but not the second, thereby teaching students only half of what they need to know (e.g., the proof of the area formula for trapezoids below gives one indication why the second case is important).

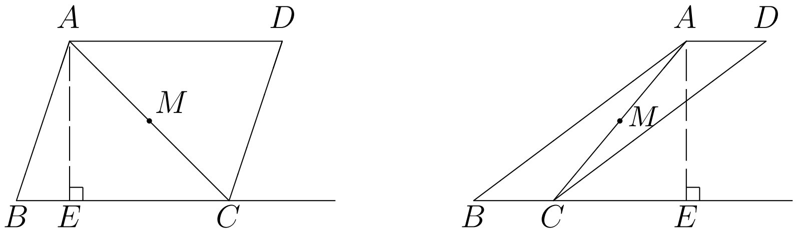
Next the area of a parallelogram $ABCD$. Drop a perpendicular from A to the opposite side BC. Call it AE. Two cases are possible, as shown below.



AE is the **height** of the parallelogram with respect to the **base** BC. From the end of §3 in Chapter 5, we know that $|AE|$ does not change if another point on AD replaces A. As before, **height** and **base** are also used to designate the *lengths* of these segments. The formula to be proved is then:

$$\textit{area of parallelogram} \quad = \quad \textit{base} \times \textit{height}$$

The proof for both cases goes as follows. Draw the diagonal AC and let M be the midpoint of AC.

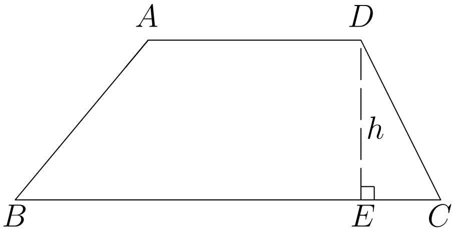


The usual argument shows that the rotation of 180° around M maps $\triangle ABC$ to $\triangle CDA$. So (M2) implies that area($\triangle ABC$) = area($\triangle CDA$). By (M3):

$$\begin{aligned}
\text{area}(ABCD) &= \text{area}(\triangle ABC) + \text{area}(\triangle CDA) \\
&= 2 \cdot \text{area}(\triangle ABC) \\
&= 2 \cdot \frac{1}{2}\,(|BC| \cdot |AE|) \\
&= |BC| \cdot |AE|
\end{aligned}$$

as desired.

We also get the formula for the area of a trapezoid $ABCD$ with $AD \parallel BC$.



Let $DE \perp BC$. Then note that $|DE|$, being the distance between the parallel lines L_{AD} and L_{BC}, is also the height of $\triangle ABD$ with respect to the base AD , and is called the **height** of the trapezoid. Again we denote the height by h. The segment AD and BC are called the **bases** of the trapezoid. We are going to prove that the area of a trapezoid is $\frac{1}{2}$ the height times the sum of bases. Precisely,

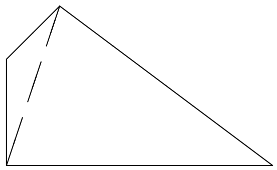
$$\mathit{area}(ABCD) \;=\; \frac{1}{2}\, h \,\{\, |AD| + |BC| \,\}$$

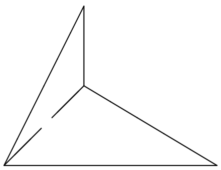
This is because:

$$\begin{aligned}
\text{area}(ABCD) &= \text{area}(\triangle BAD) + \text{area}(\triangle BDC) \\
&= \frac{1}{2}\, h \cdot |AD| + \frac{1}{2}\, h \cdot |BC| \\
&= \frac{1}{2}\, h \,\{|AD| + |BC|\},
\end{aligned}$$

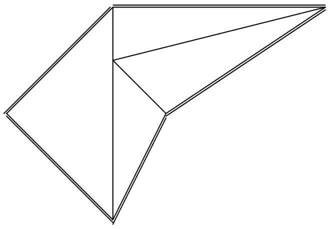
as claimed. Note that in this proof, we needed the area formula of a triangle whose altitude with respect to the given base (i.e., the height of $\triangle BDA$ with respect to the base AD) may fall outside the base (in the above picture, of course it does). This is why one must know the proof of the area formula of a triangle for this case too.

The purpose of these area formulas is not just to derive them for their own sake, although that would be entirely justified since they are answers to natural mathematical questions. However, they serve a deeper purpose. We shall show presently that triangles are the basic building blocks of polygons, and as such, the more we know about triangles the better. For example, given any quadrilateral, adding a suitable diagonal would exhibit the (inside of the) quadrilateral as the union of (the inside of) two triangles which only have a side in common but no overlap otherwise. In the following pictures, the dashed line is a diagonal of the quadrilateral. Incidentally, notice that in the figure on the right, the other diagonal would not lead to the desired result, so one knows that the choice of this diagonal cannot be made at random.



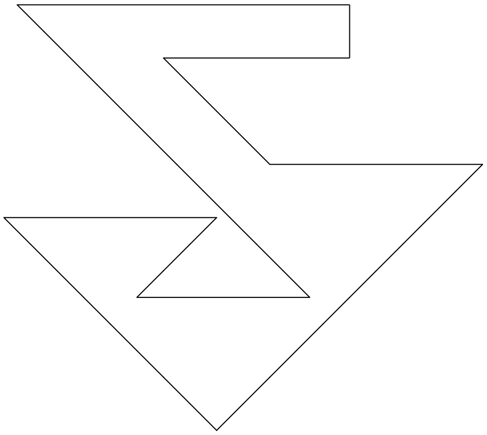


This turns out to be a universal phenomenon. To state what is know, we have to introduce a definition. As usual, the word "polygon" will be abused to mean also the polygonal region it encloses; same for "triangle". With this understood, a **triangulation** of a polygon is a union of the polygon as a finite collection of triangles $\{T_i\}$, $i = 1, 2, \ldots, k$, so that any two of these T_i's either do not intersect, or they intersect at a common vertex, or they intersect at a (complete) common edge. For example, the following is not a triangulation of the big polygon because the left triangle does not intersect any of the three triangles on the right at a complete common edge or at a common vertex:



Theorem *Every polygon has a triangulation.*

The proof is not entirely trivial. A simple example of a polygon such as the following should be enough to reveal why the proof of the Theorem has to be a complicated business.



While it is not difficult to improvise and find a way to connect the vertices of this polygon to produce a triangulation, it is not obvious, by looking at this polygon, how to describe a *general* procedure that would *always* produce a triangulation of a given polygon. Such a proof is given in Theorem 15 of Chapter 3, A. Beck, M. N. Bleicher, D. W. Crowe, *Excursions into Mathematics*, Worth Publishers, 1969, or A. K. Peters, Ltd., 2000.

Once a polygon is given a triangulation, the additivity of area (M3) implies that the area of any polygon is the sum of the areas of the triangles in its triangulation and therefore can be computed by use of any area formula of a triangle. With hindsight, the computations of the area formulas for trapezoids and parallelograms above are now seen to be nothing other than a simple application of this basic idea. In any case, the Theorem, together with the area formula of a triangle, assure us that we can compute the area of any polygon, at least in principle. This is enough for us to go on.

It remains to remark on the significance of the ability to compute the area of all polygonal regions. To this end, we have to recall the general guideline of §1 that there is little difference between the developments of the length, area, and volume functions. With this in mind, we recall the fact that, in the case of length, the ability to compute the length of all polygonal segments enabled us to compute the length of non-rectilinear curves by taking limits: we approximate a general curve by polygonal segments on the given curve, and use the lengths of these polygonal segments to approximate the length of the curve. Now polygonal regions are to area roughly what polygonal segments are to length. This is why as soon as we can compute the areas of all polygonal regions, we are free to approximate an arbitrary planar region by polygonal regions and use the areas of the latter to compute the area of the former. Therefore, in principle, we have a well-defined procedure to get an approximate value of the area of any region.[66]

We will put these ideas to use. In the special case of the disk, we get the classical formula for its area, which then turns out to finish the computation of the circumference of a circle as well (see the end of §2). But we first digress to discuss the effect of dilation on area.

Dilation and area

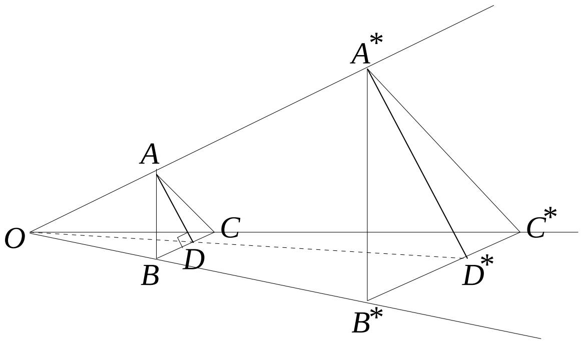
Before we give the definition of the area of a general region, we pause to discuss the effect of a dilation of ratio r on the area of a polygon. In principle, this discussion will be subsumed under a similar discussion for the case of a general region which is not necessarily a polygon. But by discussing it now, we wish to expose the elementary character of everything that is related to the area of a polygon. There is no analogous phenomenon in dimensions 3 and up.

So let $\mathcal{D}$ be a dilation with scale factor r. If $ABCD$ is a rectangle whose sides have lengths a and b, then $\mathcal{D}(ABCD)$ is also a rectangle because a dilation preserves degrees of angles (Theorem G17, Chapter 6), and a rectangle is by definition a quadrilateral with four right angles. Furthermore, the sides of $\mathcal{D}(ABCD)$ have lengths ra and rb (again Theorem G17, Chapter 6). Therefore, the area of $\mathcal{D}(ABCD)$ is $(ra)(rb) = r^2(ab)$. In other words, *if $\mathcal{D}$ is a dilation of ratio r and $\mathcal{R}$ is a rectangle, then*

$$\text{area of } \mathcal{D}(\text{R}) \;= r^2 \cdot \text{area}(\mathcal{R})$$

Suppose now ABC is a triangle with height AD and base BC. Let $\mathcal{D}(\triangle ABC) = \triangle A^*B^*C^*$, and let $\mathcal{D}(D) = D^*$. We claim A^*D^* is the height of $\triangle A^*B^*C^*$ relative to base B^*C^*. This is because dilations preserve angles and therefore $A^*D^* \perp B^*C^*$.

[66] At least those for which "area" is meaningful.



Thus if $\triangle ABC$ has height h and base b, then $\mathcal{D}(\triangle ABC)$ has height rh and base rb. Therefore

$$\text{area of } \mathcal{D}(\triangle ABC) = \frac{1}{2}\,(rh)(rb) = r^2\left(\frac{1}{2}\,ab\right).$$

In other words,

$$\text{area of } \mathcal{D}(\text{triangle}) \; = r^2 \cdot \text{area of triangle}$$

Since the area formulas of parallelograms and trapezoids were proved by using the area formula of a triangle, it follows that the areas of $\mathcal{D}$(parallelogram) and $\mathcal{D}$(trapezoid) are both r^2 times the original parallelogram and trapezoid, respectively.

In general, we compute the area of a polygon by triangulating it and apply (M3). So if a polygon $\mathcal{P}$ is triangulated into k triangles T_1, T_2, …, T_k, then $\mathcal{D}(\mathcal{P})$ is also triangulated into k triangles $\mathcal{D}(T_1)$, $\mathcal{D}(T_2)$, …, $\mathcal{D}(T_k)$, so that by (M2),

$$\begin{aligned}
\text{area}(\mathcal{D}(\mathcal{P})) &= \text{area}(\mathcal{D}(T_1)) + \text{area}(\mathcal{D}(T_2)) + \cdots + \text{area}(\mathcal{D}(T_k)) \\
&= r^2\ \text{area}(T_1) + \cdots r^2\ \text{area}(T_k) \\
&= r^2\ (\ \text{area}(T_1) + \cdots\ \text{area}(T_k)) \\
&= r^2\ \text{area}(\mathcal{P}).
\end{aligned}$$

To summarize: *if $\mathcal{D}$ is a dilation with scale factor r, then*

$$\textit{area of } \ \mathcal{D}(\textit{polygon}) = r^2 \cdot \textit{area of polygon} \qquad (13)$$

We will put this formula to good use in the next sub-section.

Note that since a similarity is the composition of a dilation and a congruence and congruence preserves area (by (M2)), formula (13) implies that *a similarity with scale* *factor r also changes area by a factor of r^2.*

Area of general regions

We next tackle the concept of area for a non-polygonal region.

Let $\mathcal{R}$ be any region.[67] Then the definition of the area of $\mathcal{R}$ will not be as elementary as that of a polygon. We shall imitate the procedure used in the definition of the length of a curve in §2 by extending what we know (i.e., area of a polygon) to what we don't know (i.e., area of a general region). Thus with $\mathcal{R}$ given, we will construct a sequence of regions P_1, P_2, $\ldots$, each of which is a union of polygons and satisfies the following condition:

> **(BC)** The boundary of each P_n gets arbitrarily close to the boundary of $\mathcal{R}$ as n gets sufficiently large.

In the most common geometric figures that come up in school mathematics, it is usually obvious how such a sequence can be constructed. We will be examining the case of the circle presently.

Following the guideline of (M4), we **define** for a general region $\mathcal{R}$:

$$\text{area}(\mathcal{R}) = \lim_{n\to\infty} \text{area}(P_n)$$

where the sequence of regions P_n satisfies condition (BC). As in the case of curves, the limit on the right is independent of the particular sequence P_1, P_2, $\ldots$ that is chosen so long as condition (BC) on the P_n's is satisfied.

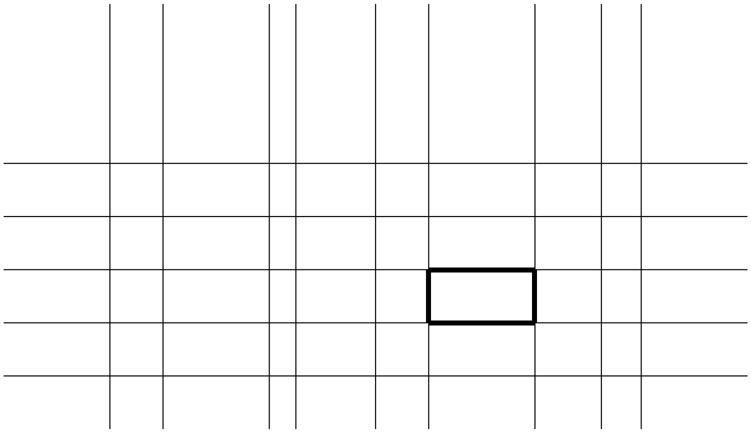
The best illustration of this definition may be that of (the region enclosed by) a circle $\mathbf{C}$. Two approximating sequences of P_n's for the circle naturally suggest themselves:

Sequence A: Inscribed polygons. Let P_3 be a regular 3-gon inscribed in $\mathbf{C}$, P_4 be a regular 4-gon inscribed in $\mathbf{C}$, and in general, let P_n be a regular n-gon

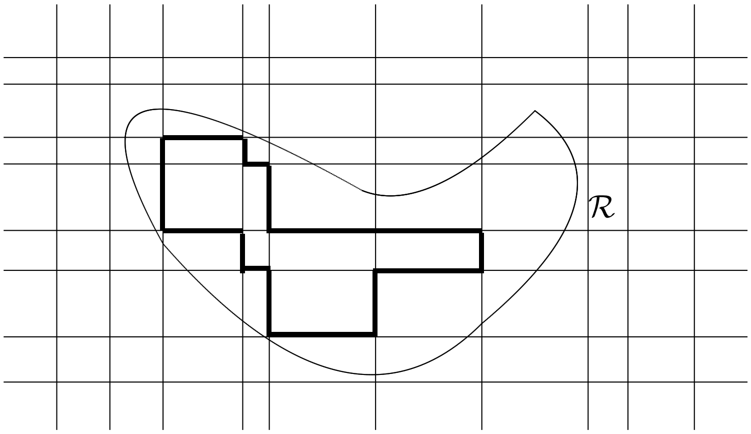
[67] Throughout this discussion, the boundary of the region is tacitly assumed to be a "nice" curve, such as a "piecewise smooth" curve. This assumption is always satisfied by the regions that come up naturally in school mathematics and, for this reason, will not be mentioned again.

inscribed in $\mathbf{C}$ for every n. (We start with P_3 because there are no polygons of 1 or 2 sides.) Then it is intuitively obvious that condition (BC) is satisfied for this sequence P_n. While this fact can be proved, we will not pursue it because it is not particularly educational.

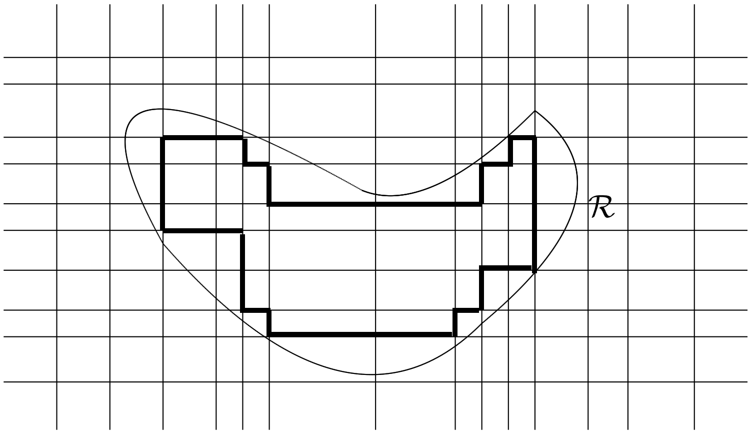
Before defining the second sequence, we need to first introduce the notion of a *grid*. Let there be a collection of horizontal lines and also a collection of vertical lines. These two collections of mutually perpendicular lines then partition the plane into rectangles whose sides are horizontal and vertical. Furthermore, these rectangles intersect only along a side, or not at all. These two collections of lines are said to form a **grid** on the plane, and the rectangles are called the **rectangles in the grid**. Here is an example of a grid and a rectangle in the grid (the thickened rectangle):



The reason we are interested in grids is due to the fact that they can be used to introduce a sequence of approximating polygons to any region. Given a region $\mathcal{R}$, we can start with a fixed grid, and define P_1 to be the union of all the rectangles in the grid that are completley inside $\mathcal{R}$. See the thickened contour below:



Next we add lines to the lattice grid by inserting one or more lines between any two existing adjacent parallel lines. This gives a new grid and P_2 is now defined to be the union of all the rectangles in the new grid which are completely inside $\mathcal{R}$. See the new thickened contour:



Similarly, we repeat the process of adding lines to the preceding grid. P_3 is by definition the union of all the rectangles in the new grid that lie completely inside $\mathcal{R}$. And so on.

We have essentially described the second sequence of regions we would use to approximate circles.

Sequence B: grid squares. Given a circle, we use a grid consisting of horizontal and vertical lines so that the distance between adjacent lines is a fixed number, say 1. Thus the smallest rectangles in the grid are all squares with sides of length 1. Let P_1 be the union of all the squares in the lattice grid lying inside the circle. Double the number of lines in the lattice grid by adding a center lines between adjacent parallel lines, and define P_2 to be the union of all the squares lying completely inside the circle. Next add center lines to the preceding grid, and define P_3 to be the union of all the squares in the new grid which lie inside the circle. And so on. We will put this sequence to use in the last sub-section of this section.

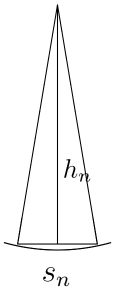
It is intuitively obvious that in either case, the sequence of polygonal regions $\{P_n\}$ can get as close to the circle $\mathbf{C}$ in the sense of condition (BC) as we please.

π and area of circles

We now specialize the considerations of the preceding sub-section to the **unit circle** $\mathbf{C}(1)$ of radius 1. Before giving the details, it may be helpful to explain what we are going to do and why we zero in on the unit circle.

We have seen in equality (12) of §2 that as soon as the length $|\mathbf{C}(1)|$ of the unit circle is known, the length of an arbitrary circle of radius r will also be known: it is r times that of the unit circle. Moreover, we shall prove (cf. (15) below) that an analogous statement is true also for area: the area of the circle of radius r is r^2 times that of the unit circle. Geometric measurements of any circle therefore come down to those of the unit circle.

We compute the area of $\mathbf{C}(1)$ by using Sequence A. Then each P_n is a regular n-gon inscribed in $\mathbf{C}(1)$. Let the length of its side be s_n and let the length of the segment from the center of $\mathbf{C}(1)$ perpendicular to a side of P_n be h_n.



Thus P_n is paved by n congruent triangles each with height h_n and base s_n, i.e., these triangles overlap only on their boundaries. Thus by (M3),

$$\begin{aligned} \text{area}(P_n) &= \underbrace{\frac{1}{2}h_n s_n + \cdots + \frac{1}{2}h_n s_n}_{n} \\ &= n\left(\frac{1}{2}h_n s_n\right) \\ &= \frac{1}{2}\; h_n(ns_n) \end{aligned}$$

Now the boundaries of P_n form a sequence of polygonal segments that approximate the unit circle $\mathbf{C}(1)$ as a curve in the sense of §2. The length of this polygonal segment, i.e., the perimeter of P_n, is $\;ns_n\;$ as there are n sides in a regular n-gon and each side has length s_n. As n *approaches infinity* (i.e., gets larger and larger), the vertices of the regular n-gon get closer and closer together so that the distances between the corners of P_n get arbitrarily small. According to definition (9) in §2, we see that the limit of (ns_n) as n approaches infinity is the length of the curve $\mathbf{C}(1)$. Moreover, as n approaches infinity, the base of the triangle gets shorter and shorter and therefore gets closer and closer to the circle $\mathbf{C}(1)$, so that the height h_n approaches the radius of $\mathbf{C}(1)$, which is 1. Therefore,

$$\begin{aligned} \text{area}(\mathbf{C}(1)) &= \lim_{n\to\infty} \;\text{area}(P_n) \\ &= \lim_{n\to\infty}\; \frac{1}{2}\,h_n(ns_n) \\ &= \frac{1}{2}\cdot 1\cdot|\mathbf{C}(1)| \end{aligned}$$

where $|\mathbf{C}(1)|$ denotes as usual the length of $\mathbf{C}(1)$. Hence,

$$\text{area}(\mathbf{C}(1)) \;=\; \frac{1}{2}\,|\mathbf{C}(1)| \tag{14}$$

At this point, we introduce the number π with the following definition:

$$\boldsymbol{\pi} \;=\; \text{area}\mathbf{(C(1)).}$$

Then (14) implies that the circumference of the unit circle is

$$|\mathbf{C}(1)| \;=\; 2\pi$$

Consider now a dilation $\mathcal{D}_r$ with scale factor r. We know from §2 that $\mathcal{D}_r(\mathbf{C}(1))$ is a circle $\mathbf{C}(r)$ with radius r and that $|\mathbf{C}(r)| = r \cdot |\mathbf{C}(1)|$ (by (12)). Hence by (14), we have obtained the well-known formula for the circumference of a circle of radius r:

$$|\mathbf{C}(r)| \;=\; 2\pi r$$

We now look into the area of $\mathcal{D}_r(\mathbf{C}(1))$. Let P_1, P_2, …be a sequence of polygons inscribed in $\mathbf{C}(1)$ satisfying condition (BC) above. Note that $\mathcal{D}_r(P_1)$, $\mathcal{D}_r(P_2)$, …is also a sequence of polygons inscribed in the circle $\mathcal{D}_r(\mathbf{C}(1))$ which satisfies condition (BC). Therefore,

$$\begin{aligned} \text{area } \mathbf{C}(1) \;&=\; \lim_{n\to\infty} \text{area}(P_n) \\ \text{area } (\mathcal{D}_r(\mathbf{C}(1))) \;&=\; \lim_{n\to\infty} \text{area}(\mathcal{D}_r(P_n)) \end{aligned}$$

However, we know from (13) in the subsection Dilation and area that

$$\text{area}(\mathcal{D}_r(P_n)) = r^2 \text{ area}(P_n).$$

Thus,

$$\begin{aligned} \text{area } (\mathcal{D}_r(\mathbf{C}(1))) \;&=\; \lim_{n\to\infty} \text{area}(\mathcal{D}_r(P_n)) \\ &=\; \lim_{n\to\infty} r^2 \text{ area}(P_n) \\ &=\; r^2 \lim_{n\to\infty} \text{area}(P_n) \\ &=\; r^2 \text{ area } \mathbf{C}(1) \\ &=\; r^2\,\pi \end{aligned}$$

But, as noted, $\mathcal{D}_r(\mathbf{C}(1)) = \mathbf{C}(r)$, the circle of radius r. Thus, we have:

$$\text{area } (\mathbf{C}(r)) \;=\; r^2 \text{ area } (\mathbf{C}(1)) \;=\; \pi\, r^2$$

We note that the same reasoning proves the more general statement that, for any region $\mathcal{R}$,

$$\text{area}(\mathcal{D}_r(\mathcal{R})) = r^2 \text{ area}(\mathcal{R}) \tag{15}$$

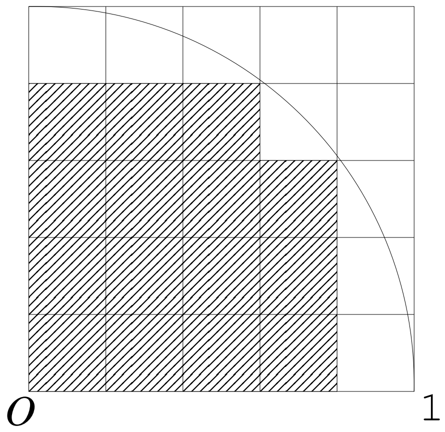
Geometric approximations of π

We will approximate the area of a circle by Sequence B. The grid specified in that sequence is now realized by the grid on a graph paper.

We start by drawing a quarter unit circle on a piece of graph paper. In principle, you should get the best graph paper possible because we are going to use the grids to directly estimate π. Now, perhaps for the first time in an honest mathematics textbook, you are going to get essential information about something other than mathematics: the grids of some of the cheap graph papers are not squares but non-square rectangles, and such a lack of accuracy will interfere with a good estimate of π. If you are the teacher and you are going to do the following hands-on activity, be prepared to spend some money to buy good graph paper.

So to simplify matters, suppose a quarter of a unit circle is drawn on a piece of graph paper so that the radius of length 1 is equal to 5 (sides of the) small squares, as shown. (Now as later, we shall use **small squares** to refer to the squares in the grid.)

The square of area 1 then contains 5^2 small squares. We want to estimate how many small square are contained in this quarter circle. The shaded polygon consists of 15 small squares in the grid. There are 7 small squares each of which is partially inside the quarter circle. Let us estimate the best we can how many small squares altogether are inside the quarter circle. Among the three small squares in the top row, a little more than 2 small squares are inside the quarter circle; let us say 2.1 small squares. By symmetry, the three small squares in the right column also contributes 2.1 small squares. As to the remaining lonely small square near the top right-hand corner, there is about 0.5 of it inside the quarter circle . Altogether the non-shaded small



squares contribute $2.1 + 2.1 + 0.5 = 4.7$ small squares, so that the total number of small squares inside the quarter circle is

$$15 + 4.7 = 19.7$$

The unit circle therefor contains about

$$4 \times 19.7 \; = \; 78.8 \quad \text{small squares}$$

Now π is the area of the unit circle, and we know that the area of 25 small squares is equal to 1. So the total area of 78.8 small squares is

$$\frac{78.8}{25} \; = \; 3.152$$

Our estimate of π is that it is roughly equal to 3.152. Taking the value of π to be 3.14159, accurate to 5 decimal digits, the percentage error of this estimate is approximately equal to

$$\frac{3.152 - 3.14159}{3.14159} \; \sim \; 0.33\%$$

While a relative error of 0.33% is very impressive, this experiment is not convincing because the percentage of guesswork needed to arrive at the final answer is too high. With a very fine and accurate grid (this is where you spend money to get good graph papers), one can reasonably get the unit 1 to be equal to anywhere between 25 to 50 small squares. Then the percentage of guesswork needed to estimate what happens to the small squares near the circle will be greatly reduced (though the counting of the total number of small squares can get dizzying).

In general, with the unit 1 equal to n small squares, then n^2 small squares have a total area of 1. If there are, after some guessing, k small squares in a quarter circle, then there are $4k$ small squares in the unit circle. Thus the area of the unit disk is

$$\pi \sim \frac{4k^2}{n^2}$$

The relative error rarely exceeds 1%.

It is recommended that all students do this activity so that they get a firm conception of what π is. Of course, this is only the beginning. As they learn more mathematics, their conception of π will broaden. Nevertheless, they need a good beginning. By contrast, most students only know "π is the ratio of circumference over diameter" when they have no idea what "circumference" means or how to go about measuring circumference accurately.

Exercises 7.3

1. A square and a rectangle have the same area, and the length of the rectangle is four times as long as the height. Which has the larger perimeter and by how much?