Lecture Notes for Math 135: Algebraic Reasoning for Teaching Mathematics

Steffen Lempp

Department of Mathematics, University of Wisconsin, Madison, WI 53706-1388, USA
E-mail address: lempp@math.wisc.edu
URL: http://www.math.wisc.edu/~lempp

Key words and phrases. algebra, equation, function, inequality, linear function, quadratic function, exponential function, student misconception, middle school mathematics

These are the lecture notes for the course Math 135:
"Algebraic Reasoning for Teaching Mathematics"
taught at UW-Madison in spring 2008.

The preparation of these lecture notes was partially supported by a faculty development grant of the College of Letters and Science and by summer support by the School of Education, both of the University of Wisconsin-Madison.

Feedback, comments, and corrections are welcome!

This text is currently in a constant state of change,
so please excuse what I am sure will be numerous errors and typos!

Thanks to Dan McGinn and my Math 135 students in spring 2008
(Rachel Burgan, Erika Calhoon, Lindsay Feest, Amanda Greisch,
Celia Hagar, Brittany Hughes, and Shannon Osgar)
for helpful comments and corrections.

© 2008 Steffen Lempp, University of Wisconsin-Madison

This material may only be reproduced and distributed for cost of reproduction and distribution for educational/non-profit purposes.
This book is currently available for free download from
`http://www.math.wisc.edu/~lempp/teach/135.pdf`.

Any dissemination of this material in modified form requires the author's written approval.

Contents

Preface vi

Introduction to the Instructor viii

Introduction to the Student x

Chapter 1. Numbers and Equations: Review and Basics 1
1.1. Arithmetic in the Whole Numbers and Beyond: A Review 1
1.1.1. Addition and Multiplication 1
1.1.2. Subtraction and Division 4
1.2. Letters and Algebraic Expressions 8
1.2.1. Letters in Algebra 8
1.2.1.1. Letters as Variables 9
1.2.1.2. Letters as Unknowns 10
1.2.1.3. Letters as Parameters 10
1.2.1.4. Letters as Constants: 11
1.2.2. Algebraic Expressions 12
1.3. Equations and Identities 13
1.4. Manipulating Algebraic Expressions 13
1.4.1. Using the definition of subtraction 14
1.4.2. "Combining Like Terms" 14
1.4.3. Removing Parentheses 15
1.5. Conventions in Algebra 15
1.5.1. Notation for Sets, and Some Special Sets of Numbers 15
1.5.2. Order of Operations and Use of Parentheses 17
1.6. Algebra as "Undoing Expressions" 19

Chapter 2. Linear Equations 21
2.1. Ratio and Proportional Reasoning 21
2.2. Velocity and Rate of Change 23
2.3. Linear Expressions 26
2.4. Line Equations 30
2.4.1. Graphing a Linear Equation 30
2.4.2. Different Forms of Equations for Lines 35

2.4.3. Varying the Parameters in a Line Equation 41
2.5. Solving a Linear Equation in One Variable 43
2.6. Solving Simultaneous Linear Equations in Two or More Variables 52

Chapter 3. Order and Linear Inequalities 59
3.1. Ordering the Numbers 59
3.2. Order and the Arithmetical Operations 61
3.3. Linear Inequalities in One Unknown 64
3.4. Estimation and Approximation 69

Chapter 4. The Concept of a Function, and Functions Closely Related to Linear Functions 72
4.1. The Concept of a Function 72
4.2. Range of a Function, Onto and One-to-One Functions, and Inverse of a Function 77
4.2.1. Range of a Function, and Onto Functions 78
4.2.2. 1–1 Functions 78
4.2.3. Inverse of a Function 79
4.3. Some Functions Closely Related to Linear Functions 82
4.3.1. Step Functions 83
4.3.2. Piecewise Linear Functions 85
4.3.3. The Absolute Value Function 86
4.3.4. Direct and Inverse Variation 89

Chapter 5. Quadratic Functions, Equations and Inequalities 92
5.1. Introduction to Quadratic Functions 92
5.2. Solving Quadratic Equations 98
5.2.1. Reviewing the Binomial Laws 98
5.2.2. Solving Quadratic Equations by Factoring 99
5.2.3. Solving Quadratic Equations by Taking the Square Root 100
5.2.4. Solving Quadratic Equations by Completing the Square 102
5.2.5. Solving Quadratic Equations by the Quadratic Formula 104
5.2.6. The number of real solutions of a quadratic equation 105
5.3. A Digression into Square Roots and the Complex Numbers 109
5.3.1. Square Roots 109
5.3.2. The Number i and the Complex Numbers 111
5.4. Graphing Quadratic Functions and Solving Quadratic Equations Graphically 113
5.5. Quadratic Inequalities 121

Chapter 6. Exponential and Logarithmic Functions, Equations and Inequalities 124
6.1. Exponentiation: A Review of the Definition 124
6.2. Motivation for Exponentiation 128
6.3. The Exponential Functions 131
6.3.1. Properties of Exponential Functions 131
6.3.2. Another Application of Exponential Functions 135
6.4. The Logarithmic Functions and Solving Exponential Equations 138
6.4.1. Definition and Properties of Logarithmic Functions 138
6.4.2. Applications of Logarithmic Functions 143
6.4.2.1. The slide rule 143
6.4.2.2. Radioactive Decay 144
6.4.2.3. Earthquakes and the Richter Scale 145
6.4.2.4. Sound Volume and Decibels 146

Bibliography 148

Index 149

Preface

"The human mind has never invented
a labor-saving device equal to algebra."
J. Willard Gibbs

The word "algebra" has several meanings in our society.

Historically, the word derives from the book "Al-Kitab al-Jabr wa-l-Muqabala" (meaning "The Compendious Book on Calculation by Completion and Balancing") written by the Persian mathematician Muhammad ibn Mūsā al-Khwārizmī (approx. 780-850 C.E.), which was the first book dealing systematically with solving linear and quadratic equations, based on earlier work by Greek and Indian mathematicians. (Curiously, the author's last name is also the source for our word "algorithm", meaning a concise list of instructions, such as in a computer program.)

In school mathematics, algebra (often called "elementary", "intermediate", "high school", or "college" algebra) follows the study of arithmetic: Whereas arithmetic deals with numbers and operations, algebra generalizes this from computing with "concrete" numbers to reasoning with "unknown" numbers ("variables", usually denoted by letters) using equations, functions, etc.

In upper-level college algebra and beyond, algebra takes on yet another meaning (which will not be the subject of this text): There, algebra (now also often called "higher algebra", "abstract algebra", or "modern algebra") abstracts from the study of numbers to the study of abstract objects which "behave like" numbers in a very broad sense; the complex numbers are a very simple (pun intended!) example of such a structure. It is the school mathematics meaning of the word "algebra" which will be the topic of this text, and in which meaning we will from now on exclusively use this word.

The history of algebra started with this meaning of the word, from the Babylonians (around 1600 B.C.E.) to the Greeks, Indians and Chinese (around 500 B.C.E. to 500 C.E.) to the Persians and Indians (around 800 C.E. to 1100 C.E.) to the Italians and French (in the 16th

century). Until the Renaissance, most of the attention in algebra centered around solving equations in one variable. How to solve quadratic equations by completing the square was already known to the Babylonians; the Italian mathematicians Scipione del Ferro and Niccolò Fontana Tartaglia independently, and the Italian mathematician Lodovico Ferrari, gave the first general solution to the cubic and quartic equation, respectively.[1] (The Italian mathematician Gerolamo Cardano was the first to publish the general solutions to cubic and quartic equations, and the formula for the cubic equation bears his name to this day.)

The French mathematician François Viète (1540 - 1603, more commonly known by his Latinized name Vieta) is credited with the first attempt at giving the modern notation for algebra we use today; before him, very cumbersome notation was used. The 18th and 19th century then saw the birth of modern algebra in the other sense mentioned above, which also led to much more general techniques for solving equations, including the proof by the Norwegian mathematician Niels Henrik Abel (in 1824) that there is no general "formula" to solve a quintic equation. (For a brief history of algebra, see the Wikipedia entry on algebra at `http://en.wikipedia.org/wiki/Algebra#History`.)

[1]Here, a *quadratic* equation is an equation involving numbers, x and x^2; a *cubic* equation is one also involving x^3; a *quartic* equation is one additionally involving x^4; and a *quintic* equation (mentioned in the following paragraph) also involves x^5.

Introduction to the Instructor

These lecture notes first review the laws of arithmetic and discuss the role of letters in algebra, and then focus on linear, quadratic and exponential equations, inequalities and functions. But rather than simply reviewing the algebra your students will have already learned in high school, these notes go beyond and study in depth the concepts underlying algebra, emphasizing the fact that there are very few basic underlying idea in algebra, which "explain" everything else there is to know about algebra: These ideas center around

- the rules of arithmetic (more precisely, the ordered field axioms), which carry over from numbers to general algebraic expressions, and
- the rules for manipulating equations and inequalities.

These basic underlying facts are contained in the few "Propositions" sprinkled throughout the lecture notes.

These notes, however, attempt to not just "cover" the material of algebra, but to put it into the right context for teaching algebra, by focusing on how real-life problems lead to algebraic problems, multiple abstract representations of the same mathematical problem, and typical student misconceptions and errors and their likely underlying causes.

These notes just provide a bare-bones guide through an algebra course for future middle school teachers. It is important to combine them with a lot of in-class discussion of the topics, and especially with actual algebra problems from school books in order to generate these discussions. For these, we recommend using the following Singapore math schoolbooks:

- Primary Math Textbook 5A (U.S. Edition),
- Primary Math Textbook 6A (U.S. Edition),
- New Elementary Math Textbook 1 (Syllabus D),
- New Elementary Math Textbook 2 (Syllabus D),
- New Elementary Math Textbook 3A (Syllabus D), and
- New Syllabus Additional Mathematics Textbook.

Only a few pages of the first and last of these books need to be used; the other four books should be used more extensively. The course web page `http://www.math.wisc.edu/~lempp/teach/135.html` gives an idea of how to integrate these various components.

We finally refer to Milgram [**2**, Chapter 8] and Wu [**6**] for similar treatments of algebra, and to Papick [**3**] and Szydlik/Koker [**5**] for very different approaches. A very careful and thorough introduction to algebra can also be found in Gross [**1**].

Introduction to the Student

These notes are designed to make you a better teacher of mathematics in middle school.

The crucial role of algebra as a stepping stone to higher mathematics has been documented numerous times, and the mathematics curricula of our nation's schools have come to reflect this by placing more and more emphasis on algebra. At the same time, it has become clear that many of our nation's middle school teachers lack much of the content knowledge to be effective teachers of mathematics and especially of algebra. (See the recent report of the National Mathematics Advisory Panel available at `http://www.ed.gov/about/bdscomm/list/mathpanel/report/final-report.pdf`, especially its chapter 4 on algebra.)

After some review of the laws of arithmetic and the order of operations, and looking more closely at role of letters in algebra, we will mainly study linear, quadratic and exponential equations, inequalities and functions. But rather than simply reviewing the algebra you have already learned in high school, we will go beyond and study in depth the concepts underlying algebra, emphasizing the following topics:

- "problem-solving": modeling real-life problems ("word problems") as mathematical problems and then interpreting the mathematical solution back into the real-world context;
- "proofs": making mathematically grounded arguments about mathematical statements and solutions;
- "analyzing student solutions": examining the rationale behind middle school and high school students' mathematical work and how it connects to prior mathematical understanding and future mathematical concepts; analyzing the strengths and weaknesses of a range of solution strategies (including standard techniques); and recognizing and identifying common student misconceptions;
- "modeling": flexible use of multiple representations such as graphs, tables, and equations (including different forms), and

how to transition back and forth between them; and using functions to model real-world phenomena;

- "symbolic proficiency": solving equations and inequalities, simplifying expressions, factoring, etc.

These lecture notes try to make the point that school algebra centers around a few basic underlying concepts which "explain" all (or at least most) of school algebra. These are

- the rules of arithmetic for numbers (including those for the ordering of numbers), and how they naturally carry over from numbers to general algebraic expressions; and
- the rules for manipulating equations and inequalities, and how they allow one to solve equations and inequalities.

We will generally label these basic facts "Propositions" to highlight their central importance. Many, if not most, of the student misconceptions and errors you will encounter in your teaching career will come from a lack of understanding of these basic rules, and how they naturally "explain" algebra.

CHAPTER 1

Numbers and Equations: Review and Basics

1.1. Arithmetic in the Whole Numbers and Beyond: A Review

1.1.1. Addition and Multiplication. Very early on in learning arithmetic, we realize that there are certain rules which make computing simpler and vastly decrease the need for memorizing "number facts". For example, we see that

$$2 + 7 = 7 + 2 \tag{1.1}$$

$$6 + 0 = 6 \tag{1.2}$$

$$3 \times 5 = 5 \times 3 \tag{1.3}$$

$$6 \times 0 = 0 \tag{1.4}$$

$$3 \times 1 = 3 \tag{1.5}$$

For a more elaborate example, we also see that

$$2 \times (3 + 9) = 2 \times 3 + 2 \times 9 \tag{1.6}$$

Is there something more general going on? Of course, you know the answer is yes, and we could write down many more such examples. However, this would be rather tiring, and we certainly cannot write down *all* such examples since there are infinitely many; so we'd rather find a better way to state these as general "rules". In ancient and medieval times, before modern algebraic notation was invented, such rules were stated in words; e.g., the general rule for (1.1) would have been stated as

"The sum of any two numbers is equal
to the sum of the two numbers in the reverse order."

Obviously, this is a rather clumsy way to state such a simple rule, and furthermore, when writing down rules this way, it is not always easy to write out the rule unambiguously. (Pause for a moment and think how you would write down the general rule for (1.6)!) And again, you know, of course, how to write down these rules in general: We use "letters" instead of numbers and think of the letters as "arbitrary"

numbers. So the above examples turn into the following general rules:

$$a + b = b + a \tag{1.7}$$

$$a + 0 = a \tag{1.8}$$

$$a \times b = b \times a \tag{1.9}$$

$$a \times 0 = 0 \tag{1.10}$$

$$a \times 1 = a \tag{1.11}$$

$$a \times (b + c) = a \times b + a \times c \tag{1.12}$$

Two related issues arise now:

(1) For which numbers a, b, and c do these rules hold?
(2) How do we know these rules are true for all numbers a, b, and c?

In elementary school, these rules can be "proved" visually for whole numbers by coming up with models for addition and multiplication. A common model for the addition of whole numbers a and b is to draw a many "objects" (e.g., apples) in a row, followed by b many objects in the same row, and then counting how many objects you have in total. Similarly, a common model for the multiplication of whole numbers a and b is to draw a rectangular "array" of a many rows of b many objects (e.g., apples), and then counting how many objects you have in total. All the above rules are now easily visualized. (E.g., see Figure 1.1 for a visual proof of (1.1) and Figure 1.2 for a visual proof of (1.3).)

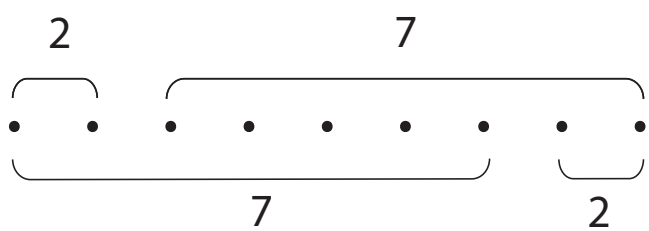

FIGURE 1.1. A visual proof that $2 + 7 = 7 + 2$

An easy extension of these arguments works for fractions (and indeed for positive real numbers) if we go from counting objects to working with points on the positive number line: On the positive number line (starting at a point identified with the number 0), mark off a "unit interval" from 0 to a new point called 1. Repeatedly marking off more unit intervals gives us points we identify with the whole numbers 2, 3, 4, etc. For a whole number $n > 0$, we can subdivide the unit interval into n many subintervals of equal length (i.e., of length $\frac{1}{n}$). Now again,

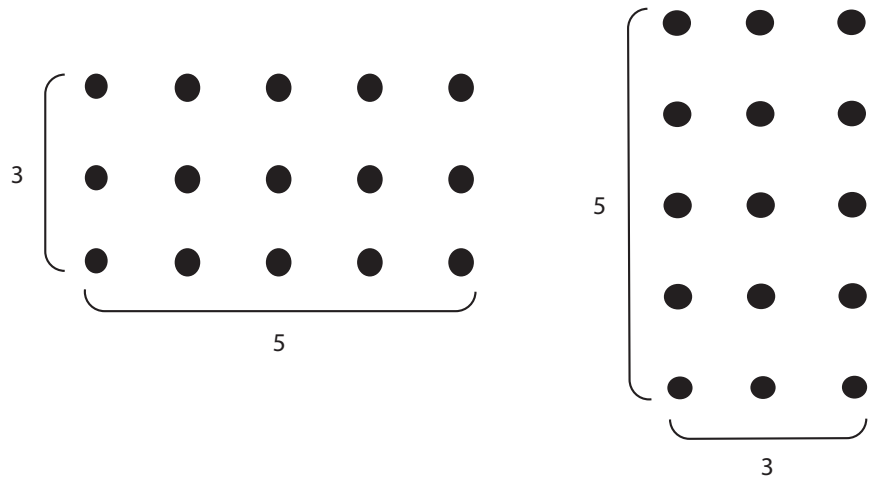

FIGURE 1.2. A visual proof that $3 \times 5 = 5 \times 3$

starting at 0, we repeatedly mark off intervals of this length and get points we identify with the fractions $\frac{1}{n}$, $\frac{2}{n}$, $\frac{3}{n}$, $\frac{4}{n}$, etc. (We also identify these points with the "directed arrows" from 0 to that point; so, e.g., $\frac{1}{2}$ is represented both by the point labeled in Figure 1.3, and by the arrow from 0 to the point labeled $\frac{1}{2}$.)

We can now model addition of fractions (and indeed for positive real numbers) a and b by marking off an interval of length a from 0, followed by an interval of length b from the point marked a. The resulting point can be identified with the sum $a + b$. Similarly, for multiplication, we can draw two positive number lines, a horizontal one going to the right and a vertical one going up, both starting at the same point. Now the rectangle bounded below by the interval on the horizontal number line from 0 to a, and bounded on the left by the interval on the vertical number line from 0 to b, has area $a \times b$ and so can be identified with this product. Once we have these models of addition and multiplication of fractions (or, more generally, positive real numbers), we can again easily verify the above rules (1.7)–(1.12). (Draw your own pictures!)

Two much more subtle rules about addition and multiplication (so subtle that you may not even think about them any more) are the following:

$$(a + b) + c = a + (b + c) \tag{1.13}$$

$$(a \times b) \times c = a \times (b \times c) \tag{1.14}$$

It is due to these rules that we can write $a + b + c$ and $a \times b \times c$ unambiguously without giving it much thought! (Again draw a picture

for each to convince yourself that they are true. For (1.14), you will have to go to a three-dimensional array.)

Exercise 1.1. Give pictures for each of (1.8) and (1.10)–(1.14) in the whole numbers.

1.1.2. Subtraction and Division. At first, students are given problems of the type $2 + 5 =$ _ or $3 \times 4 =$ __, asking them to fill in the answer. A natural extension of this is to then assign problems of the form $2 +$ _ $= 7$ or $3 \times$ _ $= 12$, asking them to fill in the missing summand or factor, respectively.

Let's look at the problem of the "missing summand" first: Solving $2 +$ _ $= 7$ corresponds, e.g., to the following word problem:

> "Ann has two apples. Her mother gives her some more apples, and now she has seven apples. How many apples did her mother give her?"

Students will quickly learn that this requires the operation of subtraction: Ann's mother gave her $7-2$, that is 5, apples. More generally, the solution to $a +$ _ $= c$ is $c - a$. But then a problem arises: What if $a > c$?[1] Obviously, you now have to go from the "objects" model to a more "abstract" model, e.g., the money model with debt: If you have \$2 and spend \$5, then you are \$3 "in debt", i.e., you have $-$\$3 dollars. (It is a good idea to emphasize the distinction in the use of the symbol $-$ in these two cases by reading $2 - 5$ as "two minus five", but -3 as "negative three".) So allowing subtraction for arbitrary numbers forces us to allow a "new" kind of numbers, the negative numbers.

So we proceed from the *whole numbers* $0, 1, 2, \ldots$ to the *integers* $0, 1, -1, 2, -2, \ldots$. You will need to allow your students to adjust to this change, since now your answer to a question like "Can I subtract a larger number from a smaller number?" will change, possibly confusing your students. Of course, one should never answer "no" to this question without some qualification; some students will know about the negative numbers at an early age (especially in Wisconsin in the winter!). But now the correct answer should be "it depends", namely, it depends on which number system we work in. At this point, you will want to extend the number line to the left also (see Figure 1.3, it can also be used to include fractions and negative fractions).

But this transition from the whole numbers to the integers immediately raises three new questions about the arithmetical operations:

[1] I can't resist at this point to tell you my favorite German math teacher joke: "What does a math teacher think if two students are in the class room and five students leave? – If three students come back, the room is empty!"

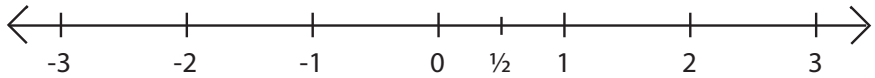

FIGURE 1.3. The number line

(1) How should we define $a + b$ and $a - b$ when one or both of a and b are negative? What about multiplication?
(2) How do we know that the rules (1.7)–(1.14) hold for addition and multiplication of arbitrary (positive or negative) numbers?
(3) Do these rules also hold for subtraction and multiplication of arbitrary (positive or negative) numbers?

Of course, you already know the answer to item (1). But how would you justify it to your students? Maybe there is another way to define addition with negative numbers? The answer is twofold: First of all, one can justify the common definition of addition of integers with a "real-life" model (such as the money-and-debt model) or with the number line (which is really just an abstraction from a real-life model). The justification for subtraction and multiplication is a bit more delicate and may not appeal to your students as much. But one can also justify the common definition of addition and multiplication of integers by the fact that we want the rules (1.7)–(1.14) to be true for addition and multiplication of integers also. So, in a sense, the rules in item (2) above really are "rules" we impose rather than rules we can "prove", unless you are satisfied with the justification for item (1) by the models for addition and multiplication.

So what about subtraction? We first define $(-b)$ as the integer which when added to b gives 0:

$$b + (-b) = 0 \tag{1.15}$$

This is now a new "rule" about addition, and $(-b)$ is called the *(additive) inverse* of b. Note that there is lots of room for confusion here: $(-b)$ (or simply $-b$ for short) is sometimes called the "negative" of b, a term I would avoid, since $-b$ need *not* be negative. In fact, $-b$ is positive if b is negative; $-b$ is 0 if b is 0; and $-b$ is negative if b is positive! (This will confuse quite a few of your students through high school and even college, so be sure to emphasize this early on.)

We can then "define" subtraction via addition: $a - b$ is defined as $a + (-b)$ for any integers a and b. Note that this works well with real-life models of subtraction: Adding a capital of -\$5 (i.e., a debt of \$5) to a capital of \$2 is the same as subtracting \$5 from \$2.

This then also gives us an answer for item (3) above: We don't have to worry about new rules for subtraction, since we can use the old rules for addition: It is fairly clear that

$$a \times (b - c) = a \times b - a \times c$$

since this is simply

$$a \times (b + (-c)) = a \times b + (-(a \times c))$$

(and since $-(a \times c) = a \times (-c)$).

On the other hand, in general $a - b = b - a$ is *false* since it would mean $a + (-b) = (-a) + b$. Similarly, $a - (b - c) = (a - b) - c$ is *false*. (How would that be rewritten?)

Let's move on to division: Again, division first comes up when a student transitions from solving $3 \times 4 =$ __ to solving $3 \times$ _ $= 12$, and solving for the missing factor corresponds, e.g., to the following word problem:

> "Ann's mother has twelve apples. If she wants to give each of her three children the same number of apples, how many will each child get?"

Again, solving this problem amounts to dividing: The solution to $3 \times$ _ $= 12$ is $12 \div 3$, or 4. More generally, the solution to $a \times$ _ $= c$ is $c \div a$. But a similar problem as for subtraction arises: Here, a may not "evenly divide" into c, e.g., in the above problem, when Ann's mother has 13 apples. Here is where it becomes really confusing to your students: There are two different ways in which school mathematics solves this problem: division with remainder, and extending to fractions. In the above example, this would mean that Ann's mother could either give four apples each to her three children and keep a "remainder" of one apple, or that she could cut one apple into thirds and give each child $4\frac{1}{3}$ apples.[2] Since both of these solutions are acceptable, you have to make it clear to your students which solution you are aiming for. (Generally, in the higher grades, you will want the second solution, but division with remainder never disappears completely from your students' mathematics: In algebra and calculus, there is division of polynomials with remainder, as we will see late in this course!)

In any case, we will for now focus on extending our numbers so that we can always divide $c \div a$ "evenly". (Unless $a = 0$: Then it

[2]There is one other model of division: "Ann's mother has 12 apples and wants to give three apples to each of her children. How many children does she have?" If you now increase the number of apples to 13, then only one kind of division makes sense since cutting up children is generally frowned upon... So be careful before you start a problem and want to vary it later!

makes no sense since it would amount to solving the original problem $0 \times _ = c$; and there will either be no solution if $c \neq 0$, or any number is a solution if $c = 0$; so $c \div 0$ *never* makes sense.) We proceed from the *whole numbers* to *fractions* (which will always be positive or 0), or more generally from the *integers* to the *rational numbers* (which may also be negative). The number line from Figure 1.3 still serves as a good visual model.

But this transition again raises a number of questions:

(1) How should we define the operations of addition, subtraction, multiplication, and division for fractions, or, more generally, for rational numbers?
(2) How do we know that the rules (1.7)–(1.15) still hold?
(3) Do these rules also hold for division?

In answering item (1), we need to be able to come up with real-life examples to justify the "correct" way to perform the arithmetical operations on fractions and more generally rational numbers. (See Exercise 1.3 below.) But again one can also justify this by wanting the rules (1.7)–(1.15) to still hold for these numbers, addressing item (2).

An elegant way to think about division is to first define the *(multiplicative) inverse* of a *nonzero* number a as that number, denoted by $\frac{1}{a}$, which when multiplied by a gives 1:

$$a \times \frac{1}{a} = 1 \tag{1.16}$$

This is now a new rule for multiplication; and it works not only when a is a positive whole number but indeed for any nonzero rational number. And again, this removes the need for new rules for division since we can reduce division to multiplication by defining $a \div b$ as $a \times \frac{1}{b}$. (See Exercise 1.4.)

We summarize all the rules about the arithmetical operations of number which we have discovered in the following

Proposition 1.2. *The following rules hold for all real numbers a, b and c:*

(commutative law of addition)	$a + b = b + a$
(associative law of addition)	$(a + b) + c = a + (b + c)$
(law of the additive identity)	$a + 0 = a$
(law of the additive inverse)	$a + (-a) = 0$

(commutative law of multiplication) $a \times b = b \times a$

(associative law of multiplication) $(a \times b) \times c = a \times (b \times c)$

(law of the multiplicative identity) $a \times 1 = a$

(law of the multiplicative inverse) $a \times \dfrac{1}{a} = 1 \qquad (\textit{when } a \neq 0)$

(distributive law) $a \times (b + c) = a \times b + a \times c$

Exercise 1.3. Give several word problems for the division $\frac{1}{2} \div \frac{1}{3}$, both for *partitive division* (as in the apple example in the main text above), and for *measurement division* (as in the footnote above).

Exercise 1.4. Explain which of the rules (1.9)–(1.12) and (1.14) carry over to division; if they don't carry over, are there corresponding rules for division?

1.2. Letters and Algebraic Expressions

1.2.1. Letters in Algebra. Arithmetic only deals with numbers and *numerical expressions*, i.e., expressions involving numbers and the arithmetical operations (addition, subtraction, multiplication, division, later on also roots, exponentiation, and logarithms). Each such expression has a definite value (or is undefined for some reason, such as division by 0 or taking the square root of a negative number). Word problems often can be reduced to evaluating numerical expressions, e.g.,

> "Ann has four apples. Her friend Mary gives her three more apples each day for a (five-day) week. How many apples does she have at the end of the week? "

Clearly, this gives rise to the numerical expression $4 + (5 \times 3)$, and evaluating this numerical expression then gives the solution

> "Ann has 19 apples at the end of the week."

One extension of this problems is to leave some numbers, e.g., the number of apples Ann has at the beginning, and the number of apples she receives each day, unspecified. In this case, we typically use letters to denote these unspecified quantities, and it is *crucial* that these letters are clearly and unambiguously specified. So let's set

$a =$ number of apples Ann has at the beginning of the week

$b =$ number of apples Ann receives from Mary each day

Now the word problem becomes

> "Ann has a apples. Her friend Mary gives her b more apples each day for a (five-day) week. How many apples does she have at the end of the week? "

This gives rise to the (more general) expression $a + (5 \times b)$, and the solution becomes

> "Ann has $a + (5 \times b)$ apples at the end of the week."

Of course, this expression can be rewritten, but we cannot simply write it as a number any more since we don't know a and b. We will cover rewriting such expressions, one of the key features of algebra, later in the course. For now, let's concentrate on the use of *letters* in algebra.

We can basically distinguish four different uses of letters:

1.2.1.1. *Letters as Variables.* This is how letters were used in the example above: Since we didn't know the number of apples Ann had at the beginning or that she received each day from Mary, we used the letters a and b for them, respectively. It doesn't really matter which letters we use, as long as we consistently use the same letter for the same quantity, use different letters for different quantities, and clearly state how each letter is used. It is often helpful to use mnemonic (i.e., easy-to-remember) letters. (E.g., in the above example, we could have used b for the number of apples Ann has at the $b$eginning of the week, and d for the number of apples Ann receives $d$aily.) Later on, we will distinguish between *independent* and *dependent* variables. In the above example, a and b are independent (since neither depends on any other quantity in the example), but if we set $c = a + (5 \times b)$ for the total number of apples at the end of the week, then c is a dependent variable, since it *depends on* a and b.

In the 5th grade Singapore math book "Primary Mathematics Textbook 5A", variables are introduced visually as bars in "bar diagrams" rather than as letters, one of the best-known features of Singapore mathematics. It is only in the next grade, in the Singapore math book "Primary Mathematics Textbook 6A", that variables are used as letters instead, and it's easy to see how one can translate the bar diagram picture into one or more equations involving variables, where the variables correspond to bars of unknown "length" in the bar diagram.

For example, in exercise 1 on page 23 of the"Primary Mathematics Textbook 5A", Ryan and Juan share \$410 between them, and Ryan receives \$100 more than Juan. The problem asks you to find Juan's share. (See Figure 1.4.) If we call Juan's share x, we can easily translate the bar diagram into an equation $(x + 100) + x = 410$.

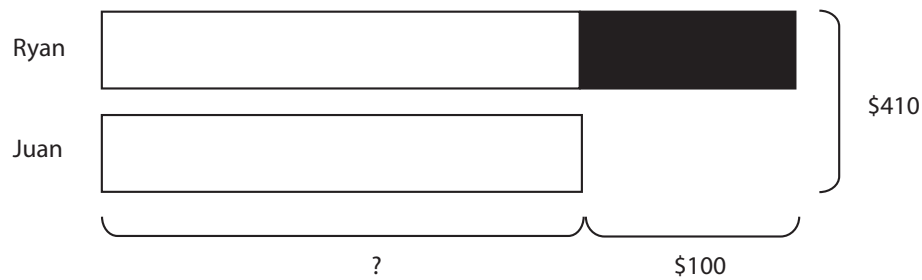

FIGURE 1.4. What is Juan's share?

1.2.1.2. *Letters as Unknowns.* This is probably the first way in which letters are introduced in school mathematics and is very similar to the previous use as variables. Now, however, we don't think of the letter as denoting an unspecified quantity, but as a specified quantity which we don't know (and usually would like to find). E.g.,

> "Ann has four apples. Her friend Mary gives her some more apples each day for a (five-day) week, the same number each day. How many apples does Mary give Ann each day of the week if Ann has 19 apples at the end of the week?"

In this case, we can express the number of apples Ann has at the end of the week in two distinct ways, as the numerical expression 19, and as the expression $4 + (5 \times b)$ where b is the number of apples she receives each day. So, to solve the problem, we need to solve the equation $4 + (5 \times b) = 19$, where the letter b denotes the unknown quantity. (For a single unknown, the letter x is usually used, but this is merely a convention.)

1.2.1.3. *Letters as Parameters.* This is probably the trickiest bit about the use of letters in algebra since the role of a letter as a parameter depends on the context. Let's look again at a variation of the above word problem

> "Ann has a apples. Her friend Mary gives her b more apples each day for a (five-day) week. At the end of the week, she has 19 apples."

This is not really a word problem yet since I haven't told you what I am asking for! In fact, I could ask for two different quantities: I could ask (1) for the number of apples Ann had at the beginning, or (2) for the number of apples Mary gave her each day. Of course, as stated, you can't give me a *numerical* answer, but you can give me an expression in each case:

For (1), you could tell me that Ann had $19-(5\times b)$ many apples at the beginning, where b is the number of apples Mary gives her each day. So in this case, we treat a as an unknown and b as a parameter, and we express the solution (i.e., in our case, a) in terms of the parameter b. Why on earth would I want to know this? Well, I can now solve the problem as soon as you tell me b, and I can solve it very quickly if you want to know a for different values of b. So, at the beginning, Ann has four apples if Mary gives her three each day, but Ann has nine apples if Mary gives her only two each day.

On the other hand, for (2), you could tell me that Mary gives Ann $\frac{19-a}{5}$ many apples each day, where a is the number of apples Ann has at the beginning. So in this case, the roles of a and b are reversed: Now a is the parameter, and b is the unknown. And the solution b is given in terms of a.

It is this switch in the roles of letters (as unknowns or parameters) which is very confusing to students, so when you reach this stage, you need to make this very clear and explicit.

Note that the distinction between a letter as a variable and a letter as a parameter is somewhat fluid, but roughly the following: A variable is a number we don't know and want to leave unspecified. A parameter is a specific number which we know but which we have not yet specified. So, in item (1) above, we can think of b as a specific number of apples Mary gives Ann every day. (Then b is a parameter.[3]) Or we can think of b as an unspecified number for which we want to study how $a = 19-(5\times b)$ (as a dependent variable) depends on the independent variable b. Thus the distinction of the role of a letter as a parameter or as a variable very much depends on the context.

1.2.1.4. *Letters as Constants:* This last use of letters is rather obvious in general: There are certain letters which are used for fixed numbers. The most common one is π, defined as the ratio of the circumference of a circle divided by its diameter. This is an irrational number and thus a non-repeating decimal, and there is no way to write it down "easily" using integers, fractions, or even roots. (The values 3.14, $\frac{22}{7}$, or 3.14159 are only approximations to its true value.) Another well-known constant in calculus is the number e, the infinite sum

$$1+\frac{1}{1}+\frac{1}{1\cdot 2}+\frac{1}{1\cdot 2\cdot 3}+\dots$$

which is again an irrational number and equals approximately 2.718. It is very useful in calculus but not really before. A third well-known

[3]E.g., b could be some really complicated number which, if plugged in now, would really make our computation very complicated.

number of the complex number i, the square root of -1. (This number is used to allow us to take the square root of negative numbers and is thus not a "real number", i.e., a number which we can write as a decimal or which we can identify with a point on the number line.) In physics, etc., there are other constants, such as g (for the gravitational constant). The only thing to remember about these constants is that one should not use the same letters for variables, unknowns or parameters when these constants are present in the context of a problem. (And the letter π should probably never ever be used in any meaning other than the usual one.)

1.2.2. Algebraic Expressions. Above, we defined a numerical expression as an expression involving numbers and the arithmetical operations (addition, subtraction, multiplication, division, later on also roots, exponentiation, and logarithms). An *algebraic expression* is an expression involving both numbers and letters as well the arithmetical operations. We think of the algebraic expression as a numerical expression where not all numbers involved are specified and thus expressed as letters.

First, a short aside: Constants take sort of an intermediate role: They should really be viewed as actual numbers, but since they are generally not expressible as rational numbers (or repeating decimals), we often leave them as letters; so the circumference of a circle of radius 2 units is normally given as 4π (the exact answer) rather than 12.57, say (which is only an approximate answer).

Algebraic expressions can be *evaluated* by replacing each letter by a number (so that the same letter is replaced by the same number everywhere in the expression, whereas different letters may, but don't have to be, replaced by different numbers). E.g., in the example of the last subsection, we evaluated $19 - (5 \times b)$ when $b = 3$ to get a value of 4, and when $b = 2$ to get a value of 9.

More importantly, algebraic expressions can be *rewritten* as algebraic expressions which take the same value no matter what numbers replace the letters in them. For this, we usually use the rules given in Proposition 1.2, which hold for all numbers and thus also for all algebraic expressions. For example, $\frac{19-a}{5}$ can be rewritten as $\frac{19}{5} - \frac{1}{5}a$ using the distributive law and the definition of subtraction and division. Learning how to rewrite algebraic expressions is one of the most important topics of algebra and requires a lot of practice.

1.3. Equations and Identities

Equality is one of the basic notions of algebra, expressing the identity of two expressions. It is first introduced in the role of "Evaluate $2 + 5 = _$". In this role, it corresponds to the way a pocket calculator would understand it, namely, "Add 2 and 5 and give the answer" rather than the role in which it is normally used in algebra, namely, simply as a statement that two expressions are equal. This can easily lead to problems: Solving $2 + _ = 7$ does *not* involve addition but subtraction, or, at first, possibly "guessing" the right answer and then adding 2 to check if the sum is 7.

What is often confusing is that there are two types of *equations*. *Identities* are equations which are always true no matter by what numbers we replace the letters, as long as the expressions on both sides are still defined. The rules in Proposition 1.2 are examples of identities, and many more can be derived from them, e.g., $a - (b + c) = a - b - c$, etc.

Any equation which is *not* an identity is thus an equation which is false when each letter is replaced by *some* number and both sides are still defined(!). So, e.g., $\frac{x}{x} = 1$ is an identity since for all nonzero x, this equation is true, and for $x = 0$, the left-hand side is undefined. On the other hand, $x - 1 = 2x - 3$ is not an identity since, e.g., when we replace x by 1, it becomes the equation $1 - 1 = 2 - 3$, which is false. Of course, if we had replaced x by 2, then we would have gotten the true equation $2 - 1 = 4 - 3$, but that is irrelevant here: The equation $x - 1 = 2x - 3$ is only an identity if it is true for *all* "values" of x, i.e., for all possible replacements of x making both sides defined.

1.4. Manipulating Algebraic Expressions

In early algebra, students are often asked to "simplify" complicated algebraic expressions, i.e., to give an algebraic expression identical to the original one using identities in each step. (Here, we call two algebraic expressions A and B *identical* if, no matter which numbers replace the letters in them, A and B evaluate to the same number.) Usually, such tasks appeal the identities from Proposition 1.2 as well as the definition of subtraction and division. But the exact way in which these steps can be justified can be obscure at times, so we will go over the two most error-prone steps in these simplifications here and justify them by our identities and definitions. Of course, I'm not trying to advocate that manipulating algebraic expressions should be taught the way I present this on the next page or so, but for you as a teacher, it should be valuable to know *why* one can perform these simplifications,

and that they are merely consequences of our identities in *arithmetic* and our definition of subtraction!

But first, let's prove a little fact about subtraction.

1.4.1. Using the definition of subtraction. In subsection 1.1.2, we defined subtraction by addition, namely, $a - b = a + (-b)$, where $(-b)$ was the number such that $b + (-b) = 0$. (Note that this definition can only be used after you have introduced negative numbers! Before then, you have to model subtraction with "real-life" situations, like taking away a part from a whole: "Mary had five apples and gave two apples to Ann. How many did she have left?")

In this subsection, we want to show that for any number a,

$$-a = (-1)a. \tag{1.17}$$

To show this, we add both $-a$ and $(-1)a$ to a:

$$a + (-a) = 0 \qquad \text{(by definition)} \tag{1.18}$$

$$\begin{aligned} a + (-1)a &= 1a + (-1)a && \text{(by (1.11))} \\ &= (1-1)a && \text{(by (1.12))} \\ &= 0a && \text{(by arithmetic)} \\ &= 0 && \text{(by (1.10))} \end{aligned} \tag{1.19}$$

Now we have to appeal to intuition (but we could also show formally using our identities!) that there is exactly one number b which, when added to a, gives 0, and so that number b equals both $-a$ and $(-1)a$ by definition and (1.18). This gives the desired identity (1.17).

1.4.2. "Combining Like Terms". Let's now start on the first of the two common steps in simplifying algebraic expressions,"combining like terms". Here is an example:

$$4x + 5y - 3x - 6y = x - y \tag{1.20}$$

Sounds pretty straightforward? It is, but when the expressions become more complicated, it can be confusing. And, most importantly, how can we justify it?

We first appeal to our identities (1.7) and (1.13):

$$4x + 5y - 3x - 6y = 4x - 3x + 5y - 6y.$$

Now we have the "like terms" adjacent and can use the distributive law:

$$4x - 3x + 5y - 6y = (4-3)x + (5-6)y = 1x + (-1)y = x - y \tag{1.21}$$

Of course, the last step uses our identities (1.8) and (1.17). But the *main step* in combining like terms consists in the use of the distributive law (1.12) in the first equality sign of (1.21)!

1.4.3. Removing Parentheses. The other common step in simplifying algebraic expressions is "removing" parentheses when adding or subtracting several terms. Here are two typical examples:

$$(a-b)+(c-d)=a-b+c-d \tag{1.22}$$

$$(a-b)-(c-d)=a-b-c+d \tag{1.23}$$

The first identity (1.22) is pretty simple to justify using the definition of subtraction and our identities (1.7) and (1.13):

$$(a-b)+(c-d)=(a+(-b))+(c+(-d))=a+(-b)+c+(-d)=a-b+c-d$$

The second identity (1.23) is one of the most common sources of errors in algebra; it requires lots of practice, and a good justification. Here it is:

$$\begin{aligned}
&(a-b)-(c-d)\\
&\quad=(a-b)+[-(c+(-d))] && \text{(by definition of subtraction)}\\
&\quad=(a-b)+[(-1)(c+(-1)d)] && \text{(by (1.17))}\\
&\quad=(a-b)+[(-1)c+(-1)^2d] && \text{(by the distributive law)}\\
&\quad=(a-b)+((-c)+1d) && \text{(by (1.17) and arithmetic)}\\
&\quad=a-b-c+d && \text{(by (1.7), (1.13) and (1.8))}
\end{aligned}$$

As you can see, this justification, even for a "small" expression like $(a-b)-(c-d)$, can become quite tricky, and so it is no wonder that removing parentheses after a minus sign is so hard for students!

1.5. Conventions in Algebra

There are a number of conventions about sets and operations which will make algebra simpler once we have mastered them, since they allow us to write statements more precisely, more succinctly, and with less clutter. (All notation and conventions introduced in this section should be introduced in school only gradually and will require repeated practice.)

1.5.1. Notation for Sets, and Some Special Sets of Numbers. One of the fundamental notion of mathematics is that of a set. Loosely speaking, a *set* is a "well-defined collection of objects". (Note that I am not actually carefully defining what a set *is* here, and for good reason: The notion of a set cannot be defined precisely, we can

only state "properties" of sets.) In any case, let's fix some notation about sets. (We will restrict ourselves to sets of numbers in the following. Not all of the notation introduced here will be used right away; so consider this subsection more of a reference for later on!)

The simplest example of a set is a finite set, and we can simply list all its elements. So the set containing exactly the numbers 0, 2, and 5 would be written as $\{0, 2, 5\}$. A special case of a finite set is the *empty set* containing no elements at all, usually written as $\{\}$ or $\emptyset$. We write $n \in S$ to denote that a number n is a member, or an *element*, of a set S.

Infinite or very large sets cannot be written this way. (For very large sets, this is at least not practical.) In that case, we will resort to using "dots" and write, e.g., $\{0, 1, \ldots, 100\}$ for the set of whole numbers less than or equal to 100. For the infinite set of all whole numbers, we can use notation like $\{0, 1, 2, \ldots\}$, etc.

Some sets of numbers are so common that one fixes standard notation: We denote

- by $\mathbf{N}$ the set of all *whole numbers* $0, 1, 2, \ldots$;
- by $\mathbf{Z}$ the set of all *integers* $0, 1, -1, 2, -2, \ldots$;
- by $\mathbf{Q}$ the set of all *rational numbers* (i.e., all numbers of the form $\frac{a}{b}$ where $a, b \in \mathbf{Z}$ and $b \neq 0$), this set is also the set of all real numbers which can be written as repeating decimals;
- by $\mathbf{R}$ the set of all *real numbers* (i.e., all numbers which can be written as decimals, possibly non-repeating); and
- by $\mathbf{C}$ the set of all *complex numbers*. (We will introduce this set in detail later on.)

(Here, $\mathbf{N}$ stands for "natural" number, which is what mathematicians usually call a whole number; $\mathbf{Z}$ stands for the German word "Zahl", meaning "number"; and $\mathbf{Q}$ stands for the word "quotient".)

Now that we have all this, we can use the "set-builder" notation $\{\ldots \mid \ldots\}$ to define new sets. E.g., the set of all positive integers can be written as $\{n \in \mathbf{Z} \mid n > 0\}$; and the set of all real numbers less than 2 can be written as $\{r \in \mathbf{R} \mid r < 2\}$.

Furthermore, later on, it will be very convenient to have notation for *intervals* of real numbers, i.e., sets of real numbers less than, greater than, or between fixed real numbers. Given real numbers a and b, we use

- $[a, b]$ for the set $\{r \in \mathbf{R} \mid a \leq r \leq b\}$;
- (a, b) for the set $\{r \in \mathbf{R} \mid a < r < b\}$;
- $[a, \infty)$ for the set $\{r \in \mathbf{R} \mid a \leq r\}$; and
- $(-\infty, b]$ for the set $\{r \in \mathbf{R} \mid r \leq b\}$.

The sets $[a, b)$, $(a, b]$, (a, ∞), and $(-\infty, b)$ are defined similarly. (Note that the symbols "∞" and "$-\infty$" do *not* denote real numbers but are merely used for convenience. Therefore, there can be no sets of the form $(a, \infty]$, $[a, \infty]$, $[-\infty, b)$, or $[-\infty, b]$.)

Finally, we will sometimes use convenient notation to combine sets. Given two sets A and B of numbers, we let

- $A \cup B = \{n \mid n \in A \text{ or } n \in B\}$ be the *union* of the sets A and B;
- $A \cap B = \{n \mid n \in A \text{ and } n \in B\}$ be the *intersection* of the sets A and B; and
- $A - B = \{n \mid n \in A \text{ and } n \notin B\}$ be the *difference* of the sets A and B.

For example, if $A = \{0, 1, 2\}$ and $B = \{0, 2, 4\}$, then

$$\begin{aligned} A \cup B &= \{0, 1, 2, 4\} \\ A \cap B &= \{0, 2\} \\ A - B &= \{1\} \end{aligned}$$

1.5.2. Order of Operations and Use of Parentheses. While the last subsection may have been rough going for you, very quickly defining a lot of notation which may be unfamiliar to you, this subsection will be fairly smooth sailing: We need to define, for numerical or algebraic expressions, in what order the operations are to be performed; and most of this will already be familiar to you from school, even though you may never have thought about it this carefully and explicitly. Note that everything we define in this subsection is only "convention", i.e., we arbitrarily define the order of operations a certain way to make things more convenient; we could have defined it completely differently, but it's certainly more convenient to have conventions everyone agrees with!

So suppose you want to write an expression denoting the sum of the difference of a and the product of 4 and b, and the difference of 2 and the quotient of 5 and c. (Note that it's hard for anyone to even comprehend what I just wrote here, so we desperately need a better way to write expressions!) In any case, we could simply write this as

$$(a - (4 \times b)) + (2 - (5 \div c)) \tag{1.24}$$

since the first rule on the order of operations is certainly that anything inside parentheses must be evaluated first; and if parentheses are nested one inside the other, then we start from the inside out.

But this expression is ridiculously redundant, of course, as you have surely noticed; we have literally cluttered it with many unnecessary

parentheses. So we use the second rule on the order of operations: All multiplications and divisions are performed before any additions and subtractions. This allows us to rewrite the above expression much more simply as

$$(a - 4 \times b) + (2 - 5 \div c) \tag{1.25}$$

We now come to the third rule on the order of operations: Additions and subtractions without parentheses are performed from left to right; and similarly multiplications and divisions are performed from left to right. This allows us to rewrite our expression even more simply as

$$a - 4 \times b + 2 - 5 \div c \tag{1.26}$$

Pause for a second here and convince yourself that we have actually used one of our rules on arithmetic to save parentheses, specifically, we have used the identity

$$(a - 4 \times b) + (2 - 5 \div c) = ((a - 4 \times b) + 2) - 5 \div c \tag{1.27}$$

Now we introduce two conventions:

- In order to avoid confusion between the multiplication symbol $\times$ and the letter x (especially when hand-written!), we drop the symbol $\times$ unless we multiply two numbers; in that latter case, we use the symbol $\cdot$ instead. So, e.g, $4 \times b$ is rewritten as $4b$, and 4×5 is rewritten as $4 \cdot 5$.
- Instead of the division symbol $\div$, we use the fraction notation. So, e.g, $5 \div b$ is rewritten as $\frac{5}{b}$, and $(5 - a) \div (2 + b)$ is rewritten as $\frac{5-a}{2+b}$. Note here that the fraction notation also saves on parentheses! If A and B are two complicated algebraic expressions, then $\frac{A}{B}$ is an abbreviated form of $(A) \div (B)$, i.e., the operations inside the expressions A and B are performed first, and then we divide the value of the former by the value of the latter.

Using these conventions, we arrive at the final simplified form of the expression (1.24) as

$$a - 4b + 2 - \frac{5}{c} \tag{1.28}$$

without any parentheses at all!

Later on, when we introduce roots and exponentiation, we also have to agree on the order of operations there. For roots, we simply indicate the order of operations by the length of the horizontal bar in the root symbol $\sqrt[n]{}$ and think of it as $(\quad)^{\frac{1}{n}}$. So, e.g., $\sqrt{2}+3$ means that we first take the square root of 2 and then add 3, whereas $\sqrt{2+3}$ means that

we first add 2 and 3 and then take the square root. For exponentiation (a fancy word for "taking the power of"), we agree that the power is taken before any other operation is performed (unless parentheses are involved). So, e.g., $a + b^2$ means that we first square b and then add a, whereas $(a + b)^2$ means that we first add a and b and then square. (A word on how to read these aloud: The first is usually read as "a plus b squared", whereas the latter is usually read as "a plus b quantity squared".)

We summarize the rules on the order of operations as follows:

(1) Any operation inside parentheses is performed before any operation outside them. If nested parentheses are involved, we perform the operations from the inside out. Any fraction $\frac{A}{B}$ (for expressions A and B) is taken to be $(A) \div (B)$. Any root $\sqrt[n]{A}$ (for an expression A) is taken to be $(A)^{\frac{1}{n}}$.
(2) If parentheses do not determine the order between operations then we first take powers, then perform any multiplications and divisions in order from left to right, and then any additions and subtractions in order from left to right.

Clearly, these conventions require quite a bit of practice before a student becomes proficient at them!

1.6. Algebra as "Undoing Expressions"

We have now set up some of the foundations of algebra: We have discussed some important properties of arithmetical operations which hold for numbers and thus must carry over to arbitrary algebraic expressions. We have isolated some of the different ways in which letters are used in algebra. We have talked about algebraic expressions and equations. We have gone over the rules which govern the evaluation of numerical expressions, and therefore *also* algebraic expressions. And we have set up some notation for sets of numbers.

Earlier, we talked about algebra as "arithmetic with letters". There is one more way in which one can look at algebra, namely, as the "art of undoing expressions". Here is an example:

> "Ann has four apples. Her friend Mary gives her three more apples each day for a (five-day) week. How many apples does she have at the end of the week?"

This is *not* an algebra but an arithmetic problem: The answer is simply given by the numerical expression $4 + (5 \cdot 3)$ apples, i.e., 19 apples. The corresponding pure arithmetic problem is $4 + (5 \cdot 3) =$ __. Now let's turn things around. Instead of *asking for* the "answer" 19,

let's *give* the answer and *ask for* one of the things in the question; let's solve $4 + (5 \cdot _) = 19$. The corresponding word problem now becomes

> "Ann has four apples. Her friend Mary gives her the same number of apples each day for a (five-day) week. At the end of the week, she has 19 apples. How many apples did Mary give Ann each day?"

Why do I call this "undoing an expression"? Let b be the number of apples Mary gives Ann each day, and c the number of apples Ann has at the end. Then we have the equation $4 + (5 \cdot b) = c$. The first problem gives b and asks for c, which amounts to simply evaluating the expression $4 + (5 \cdot b)$ when $b = 3$. The second problem gives c and asks for b. This is very different: It amounts to "solving" the equation $4 + (5 \cdot b) = 19$, i.e., "undoing" (rather than evaluating) the expression $4 + (5 \cdot b)$ to find out for which b it evaluates to 19. In this case, the "undoing" goes in two steps:

$$\begin{aligned} 4 + (5 \cdot b) &= 19 && \text{turns into} && 5 \cdot b = 19 - 4 \\ 5 \cdot b &= 19 - 4 && \text{turns into} && b = (19 - 4) \div 5 \end{aligned}$$

Now we have "undone" the expression $4+(5\cdot b)$ and reduced the solution to an arithmetic problem, namely, evaluating the numerical expression $(19 - 4) \div 5$.

Many algebra problems we will encounter later on will be of this nature. And, of course, we will encounter situations later on where it is much harder to "undo" an expression, where there may be no solution at all, or where there are several or possibly infinitely many solutions. So let's get started.

CHAPTER 2

Linear Equations

2.1. Ratio and Proportional Reasoning

Probably the first time a student encounters linearity is via the concept of *ratio*, in a problem such as this one:

> "John has twelve apples, which are either green or red. For every green apple, he has three red apples. How many red apples does he have?"

So the problem states explicitly that the ratio of the number of green apples to the number of red apples is $1 : 3$. And one can then solve the problem by finding out that John has nine red apples, so in unsimplified form, the ratio of the number of green apples to the number of red apples is $3 : 9$.

Ratio can be a confusing concept, since it can be expressed in a number of different (and "different-looking") ways:

(1) For every green apple, John has three red apples.
(2) The ratio of the number of green apples to the number of red apples is $1 : 3$.
(3) Out of every four apples, three apples are red.
(4) The ratio of the number of red apples to the number of all apples is $3 : 4$.
(5) Three quarters of John's apples are red.
(6) 75% of John's apples are red.

Here, statements (1) and (2) are basically the same, just written in different notation. However, statements (1) and (3) "look" quite different, since the former expresses a "ratio" between parts ("green apples" and "red apples"), whereas the latter expresses a ratio between a part and the whole ("red apples" and "all apples"). Statements (4), (5), and (6) formalize the ratio in statement (3) more and more abstractly, statement (4) in terms of ratio similar to statement (2); statement (5) in terms of fractions (still basically the same as (3)); and statement (6) in terms of percentage, which now takes the whole (the "unit") not as four as in all previous statements, but as 100. Figure 2.1 illustrates this ratio graphically.

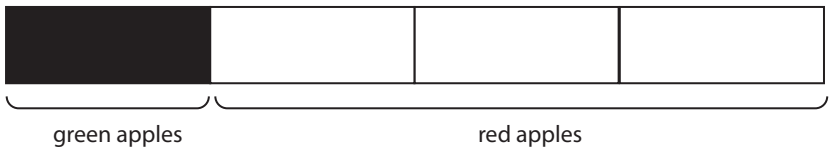

FIGURE 2.1. Ratio of the number of green apples to the number of red apples

Note that the ratio notation can be taken further by relating not just two quantities but several. E.g., let's modify the ratio problem to read:

> "John has 18 apples, which are either green, red, or yellow. For every green apple, he has three red apples and two yellow apples. How many red apples does he have?"

Now the ratio of the number of green apples to the number of red apples to the number of yellow apples is given as $1 : 3 : 2$ (and we can still solve for the number of red apples as being 9).

The concept of ratio directly leads to the concept of proportional reasoning. Let's look at the following table for the original version of our problem:

number of green apples	1	2	3	4	5	n
number of red apples	3	6	9	12	15	$3n$
number of all apples	4	8	12	16	20	$4n$

TABLE 2.1. The number of green and red apples vs. all apples

Each column contains numbers which are in ratio $1 : 3 : 4$. (The last column states this fact for a general number n of green apples.) Given any entry of the table, we can easily compute the other two entries of that column. Also note that, going from one column to the next, the numbers in each row increase by a fixed amount for each row, namely, by 1, 3, and 4, respectively. On the other hand, the ratio is usually *changed* when we add the *same* number to the entries of a column: Suppose, in our original word problem, John receives two more green apples and two more red apples, so that he now has five green apples

and eleven red apples. Then the new ratio of the number of green apples to the number of red apples is 5 : 11, which is *different* from the old ratio 1 : 3.

Similar examples can be found in many other contexts, e.g., in recipes, which call for the ingredients to be in a certain ratio (such as two cups of flour to one cup of sugar, etc.).

Percentage (from the Latin "per centum", i.e., per one hundred) is a special way to express a ratio, where one thinks of one quantity in a ratio as "the whole" and sets that number equal to 100; the other quantities are then expressed in proportion to that number. E.g., in our first word problem, the ratio of the number of green apples to the number of red apples to the number of all apples is 25 : 75 : 100, and so we can say that there are 25% green apples and 75% red apples. Usually, the "whole" is the largest of the numbers (and the others are considered "parts" of it), but this need not always be the case.

Exercise 2.1. Think of several everyday examples of ratio and percentage where the largest number is *not* considered as 100%.

2.2. Velocity and Rate of Change

Another common "linear" phenomenon is velocity:

> "A train moves at a constant velocity of 120 km/h. How far will the train travel in half an hour? In two hours? In $3\frac{1}{2}$ hours?"

Again we can make up a table:

time traveled (in hours)	0	$\frac{1}{2}$	1	2	3	$3\frac{1}{2}$
distance traveled (in km)	0	60	120	240	360	420

Table 2.2. Time and distance traveled by a train of constant velocity 120 km/h

First a brief comment: Why didn't I just call 120 km/h the *speed* of the train rather than its *velocity*? There is a small but important difference between these, and we'll get to it in more detail later on. For now, let's just remember that velocity can be negative (when the train moves backward), whereas speed is always positive (or zero, when the train is not moving). So velocity takes into account the *direction* (in one dimension, say, along a rail line), whereas speed does not.

In some sense, the above velocity problem is again a ratio problem, since time and distance traveled are in ratio 1 : 120. But that is not how

we normally think of this (even though, mathematically speaking, it is exactly the same!). The reason for this is twofold: For one, time and distance are measured in *different* units. But that alone doesn't really make it different, since, e.g., in a recipe, we might use the ratio of one cup of sugar to one teaspoon of salt. What makes time and distance "really" different is that they measure *different physical quantities*. So instead of considering this a ratio, we normally think of it as a *rate*, or more precisely, a *rate of change*, and express this rate of change as a quotient:

$$\text{velocity} = \frac{\text{distance}}{\text{time}} \tag{2.1}$$

Now Table 2.2 tells us that this quotient is fixed for any time. (However, since time and distance are measured not only in whole numbers, we don't have the property of Table 2.1 any more that as we move along a row, the numbers change by a fixed amount, since we have deliberately added some half-hour intervals.)

A quotient as in (2.1) is just one example of a rate of change. And, of course, a rate of change need not be constant at all: The train might speed up or slow down. For now, however, we will concentrate on the case of a *constant* rate of change. Examples of a constant rate of change abound in everyday situations:

(1) the cost of (loose) fruit and vegetables at the grocery store is computed by the weight (say, by the kilogram), so there is a fixed cost per kilogram;
(2) the cost of electricity you use is computed by the kWh (kilowatt hour), so there is a fixed cost per hour;
(3) the water volume of a swimming pool usually increases at a constant rate when it is filled (letting the time start when the pool starts filling up); etc.

In each case, we can identify a constant quotient. Let's make this relationship more formal: Usually, there is one independent quantity (often called x) and one dependent quantity (often called y) depending on the independent quantity. In the train example, x is the time traveled and y is the distance traveled. In the above examples (1)–(3), x is the weight, the electricity used (more precisely, the electrical energy in kWh), and the time since the pool starts filling up, respectively. And y is the cost (in both (1) and (2)) and the water volume, respectively.

Since we are assuming that the ratio (or quotient) between these two quantities x and y is constant, we'll denote this ratio by k for now;

we thus have

$$k = \frac{y}{x} \tag{2.2}$$

for any value of x and any "corresponding" y. In the train example, k is the velocity; in the other three examples, it is the cost per kilogram, the cost per kWh, and the increase in water volume (per time unit, say, per minute), respectively.

We can now rewrite (2.2) as

$$y = k \cdot x \tag{2.3}$$

to make the dependence of y on x more explicit. For our train example, switching to the more common notation of t for time, v for velocity, and s for distance, we arrive at

$$s = v \cdot t \tag{2.4}$$

The equation (2.4) is just one way to represent this situation. Another representation is via a table as in Table 2.2. A third representation is visual via a *graph* as in Figure 2.2.

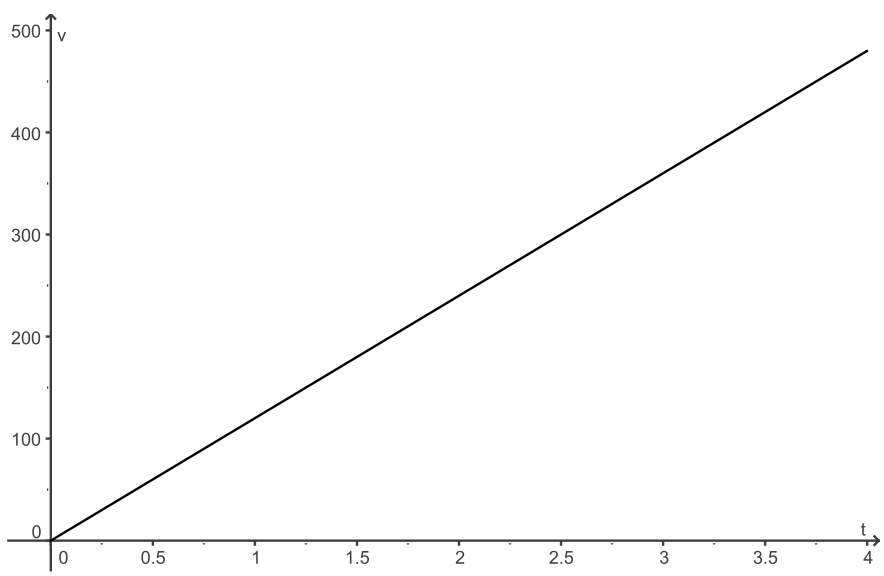

FIGURE 2.2. Time t and distance $s = v \cdot t$ traveled by a train of constant velocity $v = 120$ km/h

What are the advantages and disadvantages of these three representations?

- An equation lets us quickly compute the exact value of the distance s for any time t. As we will see later, an equation is also

handy when you need a description of how the distance varies with time inside another, more complicated computation. On the other hand, an equation is not so easy to "visualize" without some practice.

- A table lets us quickly look up the distance s for the particular values of t listed in the table. But it has us "guessing" the distance for other values of t. In fact, strictly speaking, we can't really tell the distance for other values of t at all; we are *really* just "guessing" them since we can't tell from a table how the distance "really" varies as time varies.
- A graph is a great visual aid and lets us see "at a glance" how the distance varies with time. The drawbacks are twofold: It's not really possible to read from a graph the *precise* distance at a given time t (and this is impossible for values of t "off" the graph). And it's hard to work with the graph when you have to use these values inside another, more complicated computation. (Graphs become a bit more complicated when they can no longer be drawn "without lifting the pencil" off the page, i.e., when the value of s suddenly "jumps" at some value of t. We'll talk about such graphs in more detail later.)

Since tables can only tell us the distance for finitely many values of t, they are generally not used as much in algebra. There are exceptions, however; e.g., you may know that the values of the trigonometric functions at specific multiples of π are handy to have around, and even to have memorized. (On the other hand, it's generally not hard but only tedious at worst to generate a table from an equation, say.)

From now on, we will mostly concentrate on equations and graphs. One of the main tasks in algebra will be to transfer back and forth between these two representations, a task which requires lots and lots of practice. A graphing calculator can be of great help in visualizing the graphs of more complicated equations; but for simpler equations, you, and hopefully later on your students, should be able to visualize their graphs "mentally" and sketch them quickly.

2.3. Linear Expressions

We now slightly generalize from the above examples of ratio and proportion. Here is an example:

> "A pay phone charges 50¢ for the first minute and 10¢ for each additional minute of a long-distance call."

Let's look at a table of how much a phone call costs. There are four things to note right away in Table 2.3:

number of minutes	1	2	3	4	5	6	7	8	9	10
cost in cents	50	60	70	80	90	100	110	120	130	140

TABLE 2.3. Cost of a pay phone call

(1) We still have the property as in Table 2.1 that each row increases by a fixed amount (by 1 and 10, respectively).
(2) We no longer have the properties from Tables 2.1 and 2.2 that each column has numbers in the same ratio: E.g., the ratios $1 : 50$ and $2 : 60$ from the second and third column are different.
(3) As in Table 2.1, but unlike in Table 2.2, it doesn't make "real-world" sense to look at the cost of a phone call lasting 0 minutes.
(4) As in Table 2.1, but unlike in Table 2.2, it doesn't make sense to look at the cost of a phone call lasting a number of minutes which is not a whole number (such as half a minute). We're assuming here that the phone company charges the same for a "fractional" minute (i.e., a time interval lasting less than a minute) as for a whole minute. So, e.g., a phone call lasting 2 minutes and 1 second, and another call lasting 2 minutes and 59 seconds, will both be billed as a 3-minute call by the phone company.

As you can see, there are a whole lot of new issues coming up, and we need to separate them out one by one. We'll defer the issue of whole-number minutes in item (4) until Section 4.3.1; we'll assume for now that we just "round up" and so consider only whole numbers.

The issue in item (3) is fairly easy to resolve if we just look for the "pattern" in Table 2.3: We could simply assume that the phone company charges 40¢ for a 0-minute phone call. This is, of course, not how this works out in the real world, but it makes the table follow the pattern noted in item (1).

So we are down to considering the issues from items (1) and (2). Let's look at the graph for the cost of the phone call (see Figure 2.3). Note that in light of item (3) above, I didn't draw this graph as a line but only used dots for positive whole-number values of the number of minutes.

What we would like, of course, is an equation describing this graph. Let's denote by t the number of minutes of the phone call, and by c the cost of the phone call in cents. Using the information from the original problem, we notice that $t - 1$ is the number of additional minutes, and

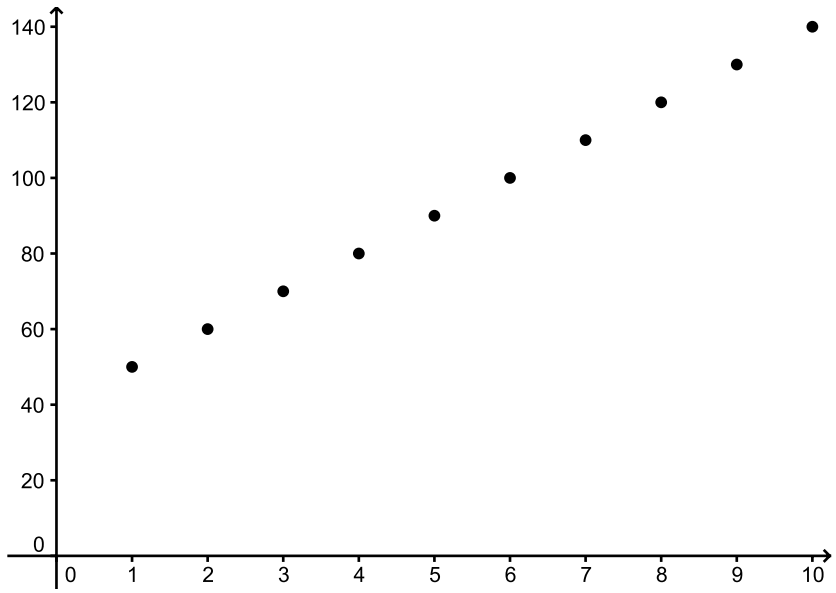

FIGURE 2.3. Cost of a pay phone call

so the cost of a phone call can be calculated by the equation

$$c = 50 + 10(t - 1) \tag{2.5}$$

where 50¢ is the cost of the first minute and 10¢ is the cost for each additional minute.

Alternatively, using the "trick" from above and letting 40¢ be the "cost" of a "0-minute" phone call, we notice that if we subtract this amount from all the other possible costs, then we do indeed get a fixed ratio, namely, 1 : 10, for all the columns in Table 2.3. This reasoning quickly leads to the equation

$$c = 10t + 40 \tag{2.6}$$

where 40¢ is the "cost" of a 0-minute call and 10¢ is the cost per "additional" minute. (It is also not hard to see that equations (2.5) and (2.6) give the same cost using our rules of arithmetic.)

It should be clear that manipulating expressions such as $10t + 40$ or $50 + 10(t - 1)$ is very important to algebra. So let's look at letters and algebraic expressions in more detail.

Here is the way Singapore math books (in Primary Mathematics 6A) first introduce letters (as variables): Think of the letter m, say, as the number of marbles in a bag (where we assume that each bag contains the same, but unknown, number of marbles). Now it makes sense

to write five bags of marbles together with three additional individual marbles as $5m + 3$ (see Figure 2.4).

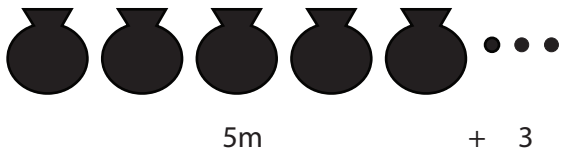

FIGURE 2.4. Representing five bags of marbles and three individual marbles

This way allows students to explore the concept of an "unknown" variable m (which may stand for any fixed number, namely, in this case, the number of marbles in each bag) while still treating m as just a "concrete" number which they can add and subtract. (Multiplication and division of variables is a bit harder to model this way and probably would require a different model.)

An expression of the form $kx+b$ (where k and b are *numbers* and x is a variable) is called a *linear expression*. (The name obviously derives from the fact, as we will study later and as you probably already know, that the graph of $y = kx + b$ is a line.) We have already seen some examples of linear expressions above. Many word problems lead to linear expressions. Here is a so-called "mixture problem":

> "A ball consisting of an iron-nickel alloy has a volume of 100cm^3. Determine its mass, given that iron has a specific density of 7.9g/cm^3 and nickel a specific density of 8.9g/cm^3."

Obviously, we can't solve this problem numerically since we are not given how much iron the alloy contains. So let's denote the volume of iron (in cm^3) in the ball by V. Then

$$7.9V + 8.9(100 - V) \tag{2.7}$$

gives the mass of the ball in grams, since the volume of the nickel must be $100 - V$ in cm^3, and so, by proportional reasoning, the mass of the iron and the nickel in the alloy is $7.9V$ and $8.9(100 - V)$ in grams, respectively.[1]

[1]Actually, the fact that iron and nickel form an alloy here changes the specific density slightly, but not enough to affect the computation enough to matter.

2.4. Line Equations

The last section should have convinced you that equations of the form

$$y = mx + b \tag{2.8}$$

and their graphs are very important in algebra; so we will study them mostly in the abstract in this section, with some asides about applications. (We have changed notation slightly here to conform with the usual conventions on these equations: We have called the independent variable x, the dependent variable y, and we have denoted the two parameters by m and b. As usual for parameters, we think of them as fixed numbers.) An equation as in (2.8) is often called a *linear equation*, but it is *not* the most general form of a linear equation as we will see later on.

2.4.1. Graphing a Linear Equation. Before we can formally define the graph of a linear equation, we need to formally define the (two-dimensional) coordinate system, which connects algebra with geometry (or, more precisely, coordinate geometry). A *coordinate system* consists of a horizontal number line (usually called the *x-axis* if x is the independent variable) and a vertical number line (usually called the *y-axis* if y is the dependent variable). (So both axes are two-way infinite lines, called the *coordinate axes*. Of course, if the variables are denoted by different letters, we also use these letters to denote the corresponding axes.) The point where the two axes intersect is called the *origin* and usually denoted by the letter O. (See Figure 2.5.)

We also introduce the set of *(ordered) pairs of reals* (x, y) by letting x and y vary over the real numbers, and by defining two pairs of reals (x_1, y_1) and (x_2, y_2) to be *equal* if both $x_1 = x_2$ and $y_1 = y_2$. We now identify each pair of reals (x, y) with the point P in the coordinate plane obtained by moving from the origin x units to the right (or, if $x < 0$, $-x$ units to the left) and y units up (or, if $y < 0$, $-y$ units down). We observe that for each pair of reals, there is exactly one such point in the coordinate plane. Similarly, for each point P in the coordinate plane, there is exactly one pair of reals found as follows: Draw a horizontal and a vertical line through the point P; let x be the number on the x-axis (viewed as a number line) at which the vertical line intersects the x-axis, and let y be the number on the y-axis (viewed as a number line) at which the horizontal line intersects the y-axis. Then the point P corresponds to the pair of reals (x, y). We call x the *x-coordinate* of P, and y the *y-coordinate* of P. Figure 2.5 illustrates this for the point $P = (2, 3)$: The vertical line through P

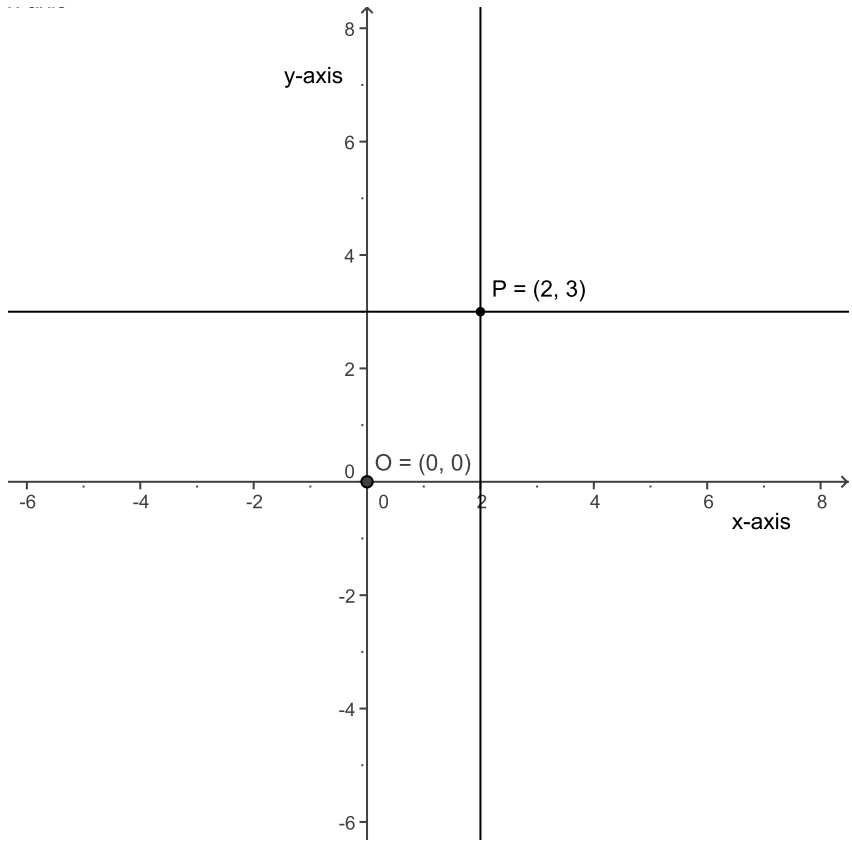

FIGURE 2.5. The coordinate system with two points labeled

intersects the x-axis at 2, and the horizontal line through P intersects the y-axis at 3.

Now the *graph* of the equation $y = mx + b$ from (2.8) consists of the set of all points (x, y) in the real plane satisfying this equation. (For an example, see Figure 2.6 for the graph of the equation $y = \frac{1}{2}x + \frac{1}{2}$.)

We want to argue that the equation (2.8) *always* describes a line in the plane. For this, we first compute the ratio of the change of y to the change of x: Fix any two distinct points $A = (x_1, y_1)$ and $B = (x_2, y_2)$ on the graph. Since A and B satisfy the equation (2.8), we have

$$y_1 = mx_1 + b \text{ and } y_2 = mx_2 + b. \tag{2.9}$$

If $x_1 = x_2$ then, by (2.9), $y_1 = y_2$ and so $A = B$, contrary to our assumption. Thus $x_1 \neq x_2$ (which will allow us below to divide by $x_2 - x_1$).

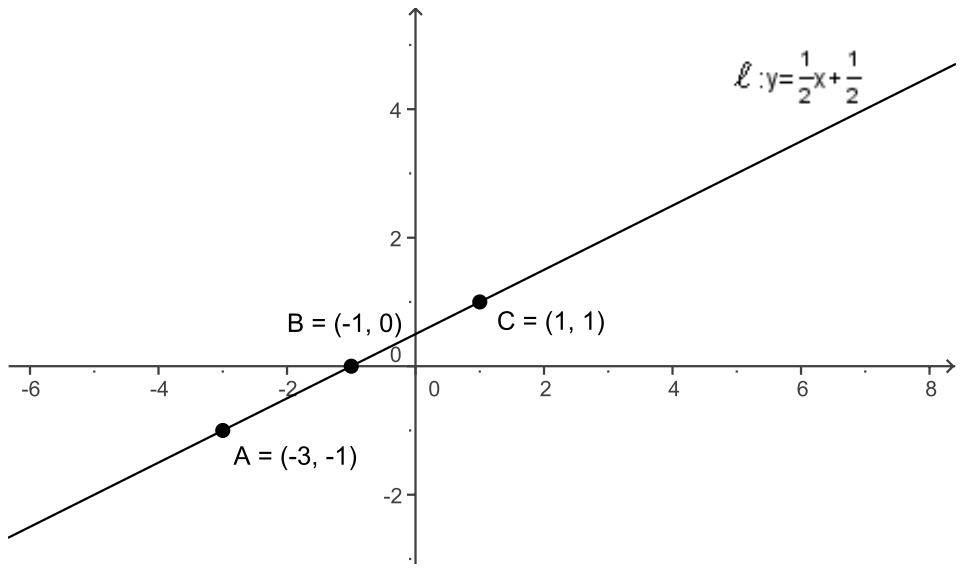

FIGURE 2.6. The line $y = \frac{1}{2}x + \frac{1}{2}$

We now compute the ratio (i.e., the quotient) of the y-change over the x-change as

$$\begin{aligned}\frac{y_2 - y_1}{x_2 - x_1} &= \frac{(mx_2 + b) - (mx_1 + b)}{x_2 - x_1} \\ &= \frac{mx_2 - mx_1}{x_2 - x_1} \\ &= \frac{m(x_2 - x_1)}{x_2 - x_1} \\ &= m\end{aligned}$$

Therefore, this ratio does not depend on the choice of the points A and B on the graph of the equation (2.8) and in fact always equals the parameter m.

Now let's pick three distinct points A, B, and B' on the graph (and assume they are arranged in this order from left to right, and that $m > 0$ so that B' and B both lie "above" A in the coordinate plane). In order to show that the graph of our equation really does form a line, we need to show that these three points are collinear. (For an example, see Figure 2.7 for the graph of the equation $y = \frac{1}{2}x + \frac{1}{2}$ with the points A, B, and B', on the graph. Don't be fooled by the picture, however! We haven't shown yet that the three points A, B and B' lie on a line! So far, we don't know that yet!) If we let the point C (C', respectively) be the point where the vertical line through B

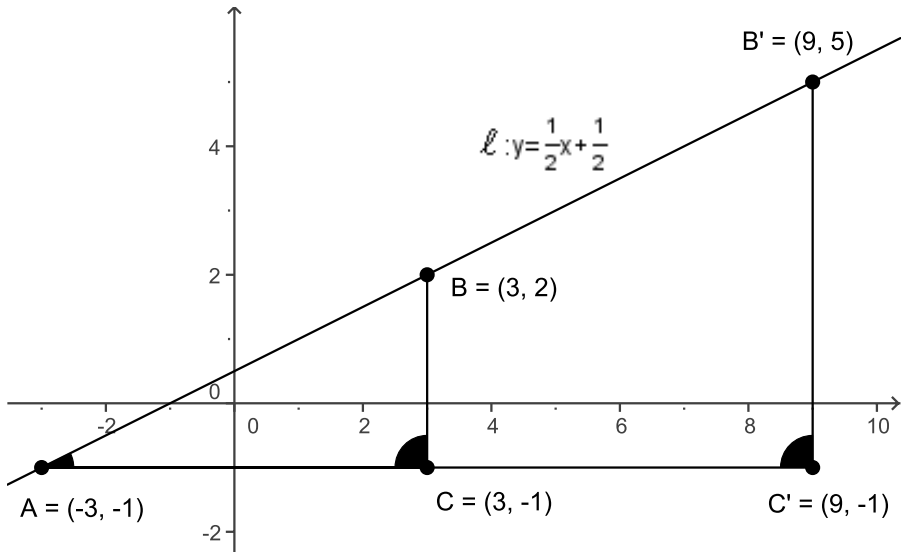

FIGURE 2.7. The slope of the line $y = \frac{1}{2}x + \frac{1}{2}$ as measured with two different triangles

(B', respectively) and the horizontal line through A intersect, then we can form two triangles $\triangle ABC$ and $\triangle AB'C'$. (Here, the picture in Figure 2.7 is correct: The points A, C, and C' really do lie on one line since the sides AC and AC' are both horizontal (i.e., parallel to the x-axis).) Now the previous calculation yields (for the lengths of the sides of the triangles) that

$$m = \frac{\overline{BC}}{\overline{CA}} = \frac{\overline{B'C'}}{\overline{C'A}}$$

Furthermore, the angles $\angle C$ and $\angle C'$ are both right angles, and thus the triangles $\triangle ABC$ and $\triangle AB'C'$ are similar. Since the sides AC and AC' are parallel (and, by similarity, the angles $\angle BAC$ and $\angle B'AC'$ coincide), the sides AB and AB' are parallel as well. Thus the points A, B, and B' on the graph of the equation (2.8) do indeed lie on a line, justifying our picture in Figure 2.7. We call the quotient m the *slope* of the line given by the equation $y = mx + b$.

EXERCISE 2.2. How does the picture in Figures 2.6 and 2.7 change when m is negative? What happens when $m = 0$? Give similar proofs that the graphs form a line in each case.

Now that we have found a visual interpretation of the parameter m, what about the parameter b? Setting $x = 0$, we obtain from the equation (2.8) that the point $(0, b)$ is on our line. This is where the line

crosses the y-axis. Therefore, b is called the *y-intercept* of the line, and the equation

$$y = mx + b$$

is called the *slope-intercept form* of the line.

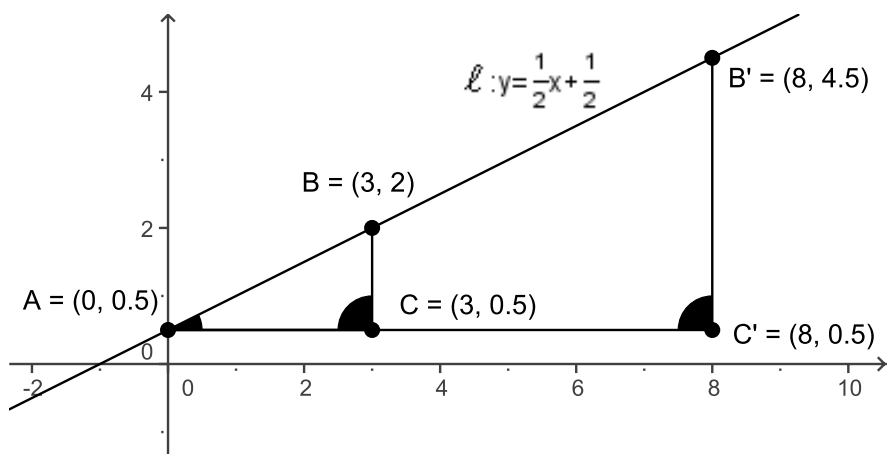

FIGURE 2.8. The slope of the line $y = \frac{1}{2}x + \frac{1}{2}$ as measured with two different triangles

We also note that the *converse* of the statement we just proved above is true: A (non-vertical) line in the coordinate plane is given by an equation of the form $y = mx + b$. (We'll talk about the special case of vertical lines a bit later.) This is because we can run the above proof "backwards": Fix a non-vertical line. It must intersect the y-axis at exactly one point $A = (0, b)$ for some real number b. Fix any other two points $B = (x_1, y_1)$ and $B' = (x_2, y_2)$ on the line. Then the line segments AB and AB' are parallel. For now, let's assume that both B and B' lie to the right of A on the line, and that the y-coordinates of both B and B' are $> b$ so that the points B and B' lie "above" A in the coordinate plane. So we can draw points C and C' (see Figure 2.8) on the intersection of the horizontal line through A, and the vertical lines through B and B', respectively. Since all three sides of the triangles $\triangle ABC$ and $\triangle AB'C'$ are pairwise parallel, these triangles are similar, and so

$$\frac{\overline{BC}}{\overline{CA}} = \frac{\overline{B'C'}}{\overline{C'A}}$$

for the lengths of their sides. This implies that the quotient $\frac{\overline{BC}}{\overline{CA}}$ does not depend on the choice of the point B, and so we denote this quotient

by m. If we now let $B = (x, y)$ be any point on the line to the right of A, then

$$m = \frac{y - b}{x - 0}$$

which can be rewritten as

$$y - b = mx$$

and so

$$y = mx + b$$

We have thus shown the following

THEOREM 2.3. *(1) The graph of any equation $y = mx + b$ is a non-vertical line with slope m intersecting the y-axis at the point $(0, b)$.*

(2) Any non-vertical line ℓ is the graph of an equation of the form $y = mx + b$ where b is determined by the y-coordinate of the point where the line ℓ intersects the y-axis, and m is the quotient

$$\frac{y_2 - y_1}{x_2 - x_1}$$

for any two distinct points $B = (x_1, y_1)$ and $B' = (x_2, y_2)$ on the line ℓ.

EXERCISE 2.4. Strictly speaking, we have shown part (2) of the above theorem only for the case when the line is "increasing", and then only in the case where the points B and B' lie to the right of the point $A = (0, b)$. Finish the proof by analyzing the cases where the line is "decreasing" or horizontal, and where one or both of the points B and B' lie to the left of the point $A = (0, b)$.

2.4.2. Different Forms of Equations for Lines. We have already seen one form of an equation for a (non-vertical) line, namely, the *slope-intercept form*

$$y = mx + b$$

where m is called the *slope*[2] and b the *y-intercept*, i.e., the y-coordinate of the point at which the line intersects the y-axis. So we can easily write down the equation of a (non-vertical) line if we are given its slope and y-intercept.

In reality, however, we are much more likely to be given two points on the line, both in geometry or in real-life applications such as the following:

[2]The Singapore math books call the slope of a line its *gradient*. In U.S. textbooks, this term is normally used in a different (but similar) meaning in calculus.

> "A car travels at constant speed. At 8:30 a.m., it crosses a state line (where the mile markers start at 0). At 10:10 a.m., it passes mile marker 100."

So letting t be the number of hours since midnight and s the distance from the state line (to be taken as negative *before* the car crosses the state line), and converting times to fractions, the problem tells us that the graph of the car's motion (which is a line since the car travels at constant speed) contains the two points $(8\frac{1}{2}, 0)$ and $(10\frac{1}{6}, 100)$.

So how do we obtain an equation for a (non-vertical) line if we only know two (distinct) points on it? The answer is, of course, that the two points tell us the slope. First remember the "mantra"

$$\text{slope} = \frac{\text{rise}}{\text{run}}$$

which really just says that the "y-change" ("rise") is proportional, i.e., in a fixed ratio, to the "x-change" ("run"). The word "rise" may be somewhat confusing since it may be negative, i.e., y may decrease as x increases; but then the slope is simply negative. (As a third case, y may just be constant as x changes, in which case the slope is 0, and the graph of the equation is a horizontal line.)

Let's call the two given points on the line P_0 and P_1, and assume that their coordinates are (x_0, y_0) and (x_1, y_1). Then the slope can be computed as

$$m = \frac{y_1 - y_0}{x_1 - x_0} \tag{2.10}$$

Similarly, given any other point $P = (x, y)$ on this line, we have

$$m = \frac{y - y_0}{x - x_0} \tag{2.11}$$

The equation (2.11) is called the *point-slope form* of a line equation and can, with after little bit of algebra, also be written as

$$y - y_0 = m(x - x_0)$$

or

$$y = mx + (-mx_0 + y_0)$$

Of course, in this last form, it is basically the same as the slope-intercept form, except that the parameter b is now more complicated.

But the point-slope form (which is useful if we are given the slope and just some point $P_0 = (x_0, y_0)$ on a line) isn't really what we were after at first, when we asked for a line equation given two (distinct) points on a (non-vertical) line. However, it is now easy to combine

equations (2.10) and (2.11) to obtain the *two-point form* of a line equation:

$$\frac{y - y_0}{x - x_0} = \frac{y_1 - y_0}{x_1 - x_0} \tag{2.12}$$

Again, with a little bit of algebra, this can be rewritten as

$$y - y_0 = \frac{y_1 - y_0}{x_1 - x_0}(x - x_0)$$

or, in slope-intercept form, as

$$y = \frac{y_1 - y_0}{x_1 - x_0}x + (-\frac{y_1 - y_0}{x_1 - x_0}x_0 + y_0)$$

And again, of course, in this last form, this is basically the same as the slope-intercept form, except that now both parameters m and b are very complicated. (Note that in equations (2.10)–(2.12), not only the letter m, but also the letters x_0, x_1, y_0, and y_1 are parameters, and only the letters x and y are variables, something which many students will find highly mystifying!)

The above three forms of the line equation all apply exactly to all non-vertical lines. There is a fourth and more general form of the line equation, which also applies to vertical lines and, not surprisingly, is called the *general form* of the line equation:

$$Ax + By + C = 0 \tag{2.13}$$

Here A, B, and C are parameters, i.e., fixed numbers, and x and y are variables, namely, the x- and y-coordinates of points in the coordinate plane.

We will *tacitly assume* that at least one of A and B is nonzero, since otherwise the equation (2.13) degenerates to the trivial equation $C = 0$, which is either true or false, independent of the values of the variables x and y, and thus does not describe a line. (The equation $C = 0$ either describes the whole plane, if the parameter C is zero, or the empty set otherwise.[3]) On the other hand, if at least one of A and B is nonzero, then the equation (2.13) really does describe a line, as we will now see.

We consider four cases for the general line equation (2.13) separately. (Here, the third case will not exclude the first two cases!)

[3]Note that in this equation $C = 0$, the letter C is the *parameter* and *not* the *unknown*! The unknowns, or variables, in $C = 0$ are x and y!

Case $A = 0$: By our tacit assumption, we must then have $B \neq 0$, and so the equation reads $By + C = 0$ and can be rewritten as

$$y = -\frac{C}{B} \tag{2.14}$$

This is just a special case of the slope-intercept form of a line, with slope 0, and so describes a horizontal line, which crosses the y-axis at the point $(0, -\frac{C}{B})$. Note that equation (2.14) does not depend on the value of the variable x, as is to be expected of a horizontal line; the equation simply fixes a particular value of the variable y, namely, $-\frac{C}{B}$.

Case $B = 0$: By our tacit assumption, we must then have $A \neq 0$, and so the equation reads $Ax + C = 0$ and can be rewritten as

$$x = -\frac{C}{A} \tag{2.15}$$

This equation does not depend on the value of the variable y and thus describes a vertical line, which crosses the x-axis at the point $(-\frac{C}{A}, 0)$; the equation simply fixes a particular value of the variable x, namely, $-\frac{C}{A}$. Note that equation (2.15) *cannot* be rewritten in slope-intercept form or either of the other two forms introduced earlier in this section.

Case $C = 0$: First, note that in this case, one of A or B may also equal 0, so this case does not exclude either of the first two special cases. However, since the equation now reads $Ax + By = 0$, we immediately see that the origin $(0, 0)$ is a point on the line. If A or B also equals 0, then, by the previous cases, the equation describes the x-axis or the y-axis as a line, respectively. If neither A nor B equals 0, then the equation can be rewritten as

$$y = -\frac{A}{B}\,x \tag{2.16}$$

Since this line crosses both the x-axis and the y-axis at the origin, both the y-intercept (defined before as the y-coordinate of the point where a line crosses the y-axis) and the x-intercept (defined now as the x-coordinate of the point where a line crosses the x-axis) equals 0. And the slope of the line clearly equals $m = -\frac{A}{B}$.

Case $A, B, C \neq 0$: In that case, the equation (2.13) can be rewritten as

$$y = -\frac{A}{B}\,x - \frac{C}{B}$$

and so has nonzero slope $m = -\frac{A}{B}$. Also, since the parameter C does not equal 0, the line does not cross the axes at the origin. Rather, by

setting $x = 0$ and and solving for y, and then setting $y = 0$ and solving for x, respectively, it is not hard to see that the line contains the two (distinct) points $(-\frac{C}{A}, 0)$ and $(0, -\frac{C}{B})$ on the x- and y-axis, respectively. So the x-intercept of the line is $-\frac{C}{B}$, and its y-intercept is $-\frac{C}{A}$. Since we can multiply the equation (2.13) by any nonzero number without affecting its truth, we may assume (having multiplied equation (2.13) by the appropriate number) that the parameter C (after "renaming" of parameters) equals -1. Then equation (2.13) reads (with the *new* values for our parameters A, B, and C)

$$Ax + By - 1 = 0$$

or

$$Ax + By = 1$$

Now the x- and y-intercepts are $\frac{1}{A}$ and $\frac{1}{B}$, respectively. Defining new parameters a and b to equal $\frac{1}{A}$ and $\frac{1}{B}$, respectively, we can thus rewrite equation (2.13) (still assuming that the parameter C equals -1) as

$$\frac{x}{a} + \frac{y}{b} = 1 \tag{2.17}$$

This is another "normal form" of an equation of a line (*not* containing the origin). It is sometimes called the *intercept form* of such a line. (See Figure 2.9 for the example $\frac{x}{-2} + \frac{y}{1} = 1$.)

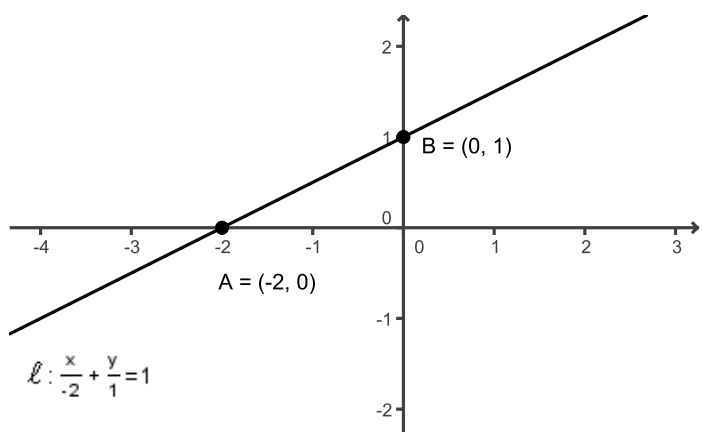

FIGURE 2.9. The line $\frac{x}{-2} + \frac{y}{1} = 1$ with x-intercept $a = -2$ and y-intercept $b = 1$

Finally, let's look at the five different forms of a line equation in terms of a "real-life" example. In each case, a car will cross a state line (where the mile markers start at 0) and travel at a constant speed. We use the independent variable t for the time elapsed in hours since

noon, and the dependent variable s for the distance traveled in miles starting at the state line (taking s to be negative before the car crosses the state line). In all versions of this example, we will get an equation for the same line, see Figure 2.10.

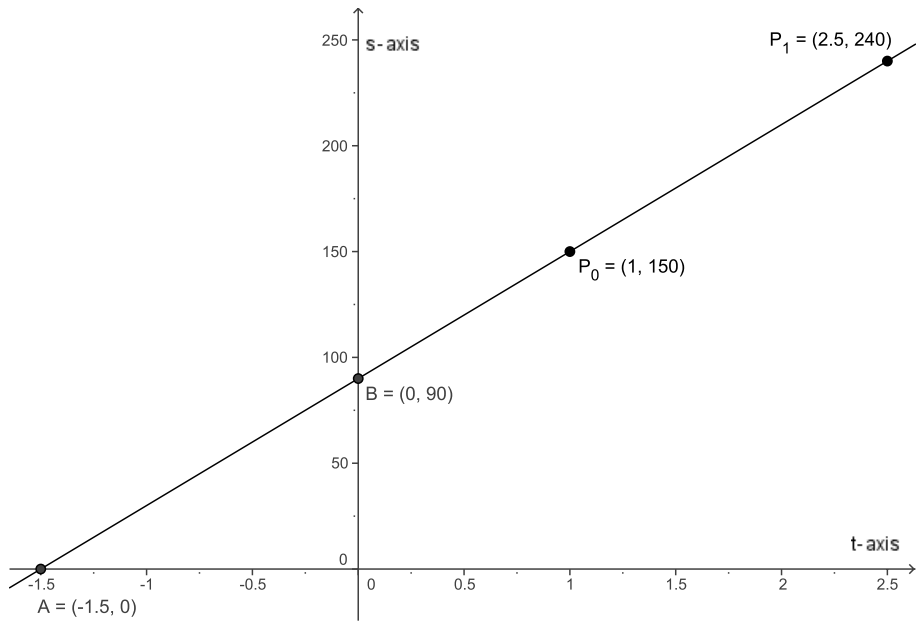

FIGURE 2.10. The line $s = 60t+90$ describing the travel of a car

Slope-intercept form:

> "A car travels at a constant speed of 60mph. At noon, it has traveled 90 miles past the state line."
> Slope $m = 60$, s-intercept $b = 90$, equation
> $$s = 60t + 90$$

Point-slope form:

> "A car travels at a constant speed of 60mph. At 1 p.m., it has traveled 150 miles past the state line."
> Slope $m = 60$, point $P_0 = (1, 150)$, equation
> $$60 = \frac{s - 150}{t - 1} \text{ or } s - 150 = 60(t - 1)$$

Two-point form:

> "A car travels at a constant speed. At 1 p.m., it has traveled 150 miles past the state line; at 2:30 p.m., it

has traveled 240 miles past the state line."
Points $P_0 = (1, 150)$ and $P_1 = (2.5, 240)$, equation

$$\frac{s-150}{t-1} = \frac{240-150}{2.5-1} \text{ or } s - 150 = \frac{240-150}{2.5-1}(t-1)$$

Intercept form and general form:

"A car travels at a constant speed. At 10:30 a.m., it crosses the state line; at noon, it has traveled 90 miles past the state line."
t-intercept $a = -1.5$, s-intercept $b = 90$, equation in intercept form

$$\frac{t}{-1.5} + \frac{s}{90} = 1$$

and (after clearing denominators by multiplying by -90) in general form

$$60t - s + 90 = 0$$

Although the different phrasings of this problem may seem somewhat contrived, they show that each form of a line equation can be modeled by a "real-life" example, which is what I was trying to convince you of.

Exercise 2.5. Find a "very different" real-life example for the equation $y = -2x - 2$ and explain how this problem would be phrased in terms of the four different forms of the line equation (slope-intercept, point-slope, two-point, and intercept form).

2.4.3. Varying the Parameters in a Line Equation. Let's now see what happens when we vary the parameters. Most of this should be clear from the "visual" or "physical" interpretation of the parameters in the different forms of the line equation covered in the previous subsection.

In the slope-intercept form of a line equation

$$y = mx + b \tag{2.18}$$

what happens if we vary the slope m and the y-intercept b? (See Figure 2.11 for examples.)

Here are some fairly obvious remarks about the slope m:

- If m is positive, then the line described by (2.18) "goes up" or "is increasing", i.e., y increases as x increases.
- If m is negative, then the line described by (2.18) "goes down" or "is decreasing", i.e., y decreases as x increases.

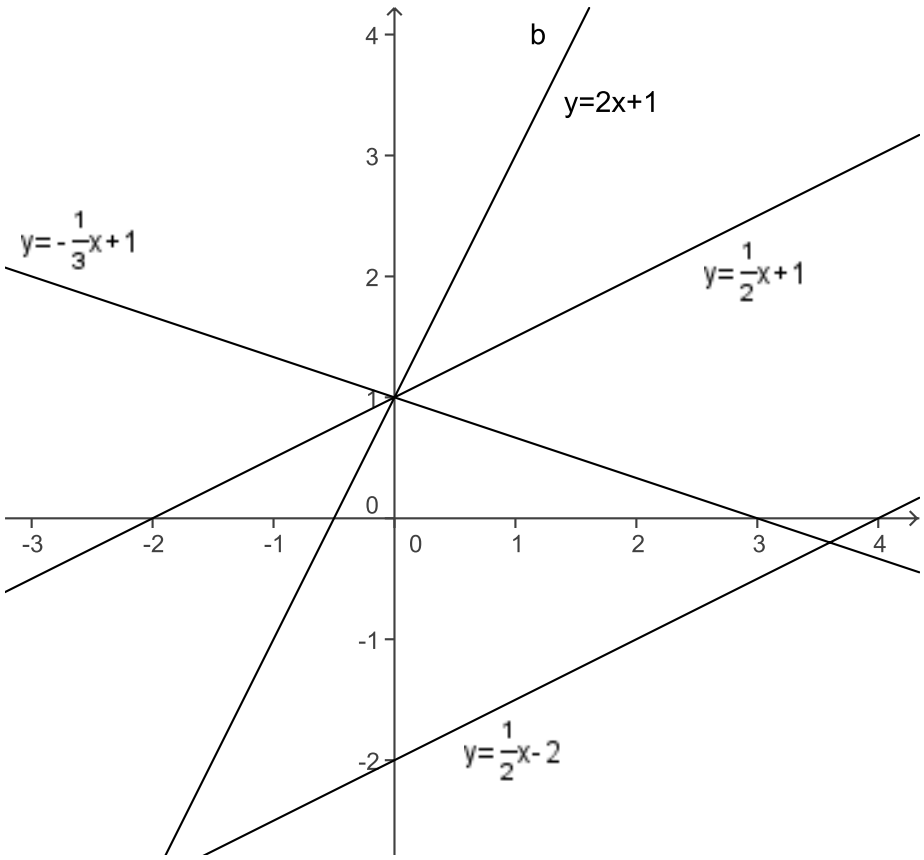

FIGURE 2.11. Examples of varying slope and y-intercept of a line

- If m is 0, then the line described by (2.18) is horizontal, i.e., y remains constant as x varies.
- As m increases (and assuming it is always positive), the line described by (2.18) becomes "steeper".
- As m decreases (and assuming it is always positive), the line described by (2.18) becomes "flatter" or "less steep".

Similarly, here are some remarks about the y-intercept b:

- If b is positive, then the line described by (2.18) crosses the y-axis "above" the origin.
- If b is negative, then the line described by (2.18) crosses the y-axis "below" the origin.
- If b is 0, then the line described by (2.18) crosses the y-axis at the origin.

- As b increases or decreases, the point at which the line described by (2.18) crosses the y-axis moves "up" or "down", respectively.

Similar comments can be made about the x- and y-intercepts in the intercept form of the line equation

$$\frac{x}{a} + \frac{y}{b} = 1$$

2.5. Solving a Linear Equation in One Variable

After all this long preparation, let's now "solve" linear equations! We'll again start with two examples:

> "A pool is being filled with water at a constant rate of 50 ℓ/min (liters per minute), starting at noon. If the pool already contained 1,000 liters of water at noon, how much water will be in the pool after 30 minutes?"

> "A pool is being filled with water at a constant rate of 50 ℓ/min, starting at noon. If the pool already contained 1,000 liters of water at noon, what time will it be when the pool contains 2,500 liters?"

We can see that the time t (in minutes since noon) and the water volume V in the pool (in liters) are related in both examples by the equation

$$V = 50t + 1000 \tag{2.19}$$

since the "slope" is the rate of 50 ℓ/min, and the "V-intercept" is the 1,000 liters of water already in the pool at noon. Now the first example above is just an arithmetic problem: We replace t by 30 and obtain the answer $V = 50 \cdot 30 + 1000 = 2500$ (in liters). On the other hand, the second problem gives $V = 2500$ and asks for a solution of the equation

$$50t + 1000 = 2500 \tag{2.20}$$

This requires "undoing" the operation of "computing the volume from the time" as explained in section 1.6, i.e., it requires rewriting the equation by isolating t. We undo (2.20) in two steps:

$$\begin{array}{lll} 50t + 1000 = 2500 & \text{turns into} & 50t = 1500, \text{ and} \\ 50t \qquad\quad = 1500 & \text{turns into} & t = 30 \end{array}$$

This then gives the answer: "It will be 12:30 p.m. when the pool contains 2,500 liters."

More precisely, what we are saying is that the following three equations are *equivalent* (namely, have the same solutions in t):

$$50t + 1000 = 2500 \tag{2.21}$$

$$50t = 1500 \tag{2.22}$$

$$t = 30 \tag{2.23}$$

But *why* do we know that they are equivalent? Notice that we really have to show two things:

(1) Any solution to equation (2.21) is also a solution to equation (2.23), *and*
(2) any solution to equation (2.23) is also a solution to equation (2.21).

Let's address item (1) first: Equation (2.22) is obtained from equation (2.21) by subtracting 1000 "on both sides". But, if we replace t in equation (2.21) by any number which makes both sides of the equation equal, then this equality remains true if we subtract 1000 from both sides. (After all, the two sides of equation (2.21) represent the same number before the subtraction, so they must still represent the same number after the subtraction!) The same is true in going from equation (2.22) to equation (2.23): If we replace t in equation (2.22) by any number which makes both sides of the equation equal, then this equality remains true after we divide by 50 on both sides.

The argument for item (2) is similar, but with a crucial difference: We first have to show that if we replace t in equation (2.23) by any number which makes both sides of the equation equal, then this equality is true *before* we divide both sides by 50. There is only one way to show this: Note that equation (2.22) can be obtained from equation (2.23) by *multiplying* both sides by 50! Now we can argue as for item (1): If we replace t in equation (2.23) by any number which makes both sides of the equation equal, then this equality remains true if we *multiply* both sides by 50, and the resulting equation is equation (2.22). A similar argument shows that any solution to equation (2.22) is a solution to equation (2.21), since equation (2.21) can be obtained from equation (2.22) by *adding* 1000 to both sides of the equation. Both steps of the argument for item (2) can be viewed as "undoing" the corresponding steps of the argument for item (1).

This leads us to two crucial observations about "solving equations" in general:

Proposition 2.6. *(1) Any solution to an equation $A = B$ (where A and B are algebraic expressions) remains a solution*

if we add or subtract the same algebraic expression on both sides of the equation; and vice versa, any solution to the equation after *adding or subtracting the same algebraic expression is also a solution of the original equation $A = B$.*

(2) *Any solution to an equation $A = B$ (where A and B are algebraic expressions) remains a solution if we multiply or divide both sides of the equation by the same* nonzero *algebraic expression; and vice versa, any solution to the equation* after *multiplying or dividing by the same nonzero algebraic expression is also a solution of the original equation $A = B$. (Here, an algebraic expression is* nonzero *if it cannot equal 0 no matter what numbers replace the letters in the expression.)*

Here are two fine points which you can ignore on a first reading, but which are important in general:

Remark 2.7. *There is small caveat here: If the algebraic expression C in Proposition 2.6, which we add, subtract, multiply or divide by on both sides of an equation cannot be computed for all values of its variables, then our proposition only holds for those values at which C can be computed.*

There is also an extension of Proposition 2.6 (2): If the algebraic expression C equals 0 for some but not all replacements of letters in C by numbers, then Proposition 2.6 (2) remains true for those values at which C does not equal 0.

Let's first look at the caveat. Here are two very simple examples: (1) If we add $\frac{1}{x}$ to both sides of the equation $x = 0$, say, then the new equation $x+\frac{1}{x} = \frac{1}{x}$ obviously does *not* have the same solutions as $x = 0$; but the proposition still holds for all values of x at which $\frac{1}{x}$ is defined, namely, whenever x is nonzero: Both equations $x = 0$ and $x + \frac{1}{x} = \frac{1}{x}$ are false for all nonzero x, as the proposition states. (2) If we add $\sqrt{x}$ to both sides of the equation $x = -1$, say, then the new equation $x + \sqrt{x} = -1 + \sqrt{x}$ obviously does *not* have the same solutions as $x = -1$; but the proposition still holds for all values of x at which $\sqrt{x}$ is defined, namely, whenever x is non-negative: Both $x = -1$ and $x + \sqrt{x} = -1 + \sqrt{x}$ are false for all non-negative x as the proposition states.

As for the extension of Proposition 2.6 (2), here is a very simple example: The equation $x = 1$ obviously does not have the same solutions as the equations $x^2 = x$ (obtained by multiplying both sides by x). However, if we only look at the solutions to both equations where x

is nonzero, then they *do* indeed have the same solutions, namely, only $x = 1$.

Now for the proof of Proposition 2.6: Adding to, subtracting from, multiplying by a nonzero expression, or dividing by a nonzero expression, both sides of an equation "preserves" all solutions of the original equation; and since these steps can be reversed by subtracting, adding, dividing, and multiplying, respectively, we also don't get any extra solutions.

Notice that the crucial part in Proposition 2.6 is the "vice versa" part: Not only do we "preserve" a solution by adding, subtracting, multiplying or dividing both sides of an equation; we also don't get any *new* solution. Of course, in part (2) of Proposition 2.6, we definitely need the restriction that we don't multiply or divide by 0 (or an expression which could equal 0): Dividing by 0 is not allowed anyway, and multiplying an equation $A = B$ by 0 leads to the equation $0 = 0$, which almost always has *more* solutions than the original equation, namely, any number is a "solution" to $0 = 0$. A more sophisticated example (and an example of a very common error) is to divide both sides of the equation $x^2 = x$ by x to "arrive" at the equation $x = 1$. Clearly, x may be zero, so this would not be allowed under Proposition 2.6, and it is also easy to see that Proposition 2.6 fails in this example: The equation $x^2 = x$ has *two solutions*, namely, 0 and 1, whereas the equation $x = 1$ only has the solution 1. (However, as noted above, the extension of our proposition still allows us to state that the *nonzero* solutions of $x^2 = x$ and $x = 1$ agree!)

More generally, why am I putting so much emphasis on the "vice versa" part? After all, it's obvious that you shouldn't multiply both sides of an equation by 0 while solving an equation. The reason is that this "vice versa" part will become much more subtle later on and leads to frequent mistakes. For example, if we *square* an equation $A = B$, then any solution to the original equation $A = B$ is still a solution to the equation $A^2 = B^2$ *after* squaring, but *not* vice versa, as you can see from the simple example $x = 1$: It has only one solution, namely, 1; but the equation $x^2 = 1$ (i.e., after squaring) has two solutions, namely, 1 and -1. So as we extend Proposition 2.6 later on (e.g., in Proposition 5.3) to allow more "operations" on equations, we always have to make sure that the "vice versa" part also holds, since it can fail in subtle ways!

Let's now vary the pool problem a bit to arrive at the general form of a linear equation:

> "A pool is being filled at a constant rate of 50 ℓ/min, starting at noon, at which time it already contains 1,000 liters of water. If that same pool had contained 1,600 liters of water at noon and had been filled at the rate of only 30 ℓ/min, starting at noon, it would have been full at the same time. At what time will the pool be full? How much water can the pool hold?"

If we still let t be the time (in minutes) since noon, the expression $50t + 1000$ describes the water volume under the first scenario, while the expression $30t + 1600$ describes it under the second scenario. The problem tells us that the pool is full (and so contains the same volume of water) at the same time t, so we arrive at the equation

$$50t + 1000 = 30t + 1600 \tag{2.24}$$

Now we can use Proposition 2.6 three times to "isolate" t: Equation (2.24) is equivalent to

$$20t + 1000 = 1600$$

by subtracting $30t$ from both sides. Next we subtract 1000 from both sides to arrive at the equivalent equation

$$20t = 600$$

Finally, we divide by 20 on both sides to arrive at the solution

$$t = 30$$

which yields the answer "The pool will full at 12:30 p.m." A small further computation shows that the pool can hold 2,500 liters (since $50 \cdot 30 + 1000 = 30 \cdot 30 + 1600 = 2500$).

There is one other way to solve our pool problems, namely, graphically. Let's return to the first equation we tried to solve:

$$50t + 1000 = 2500 \tag{2.20}$$

Consider the equation describing the water volume in terms of the time:

$$V = 50t + 1000 \tag{2.19}$$

We can view solving (2.20) as determining the t-coordinate of the point $P = (30, 2500)$ where the line described by equation (2.19) intersects the horizontal line $V = 2500$. (See Figure 2.12.)

Let's turn to the other equation we solved in this section:

$$50t + 1000 = 30t + 1600 \tag{2.24}$$

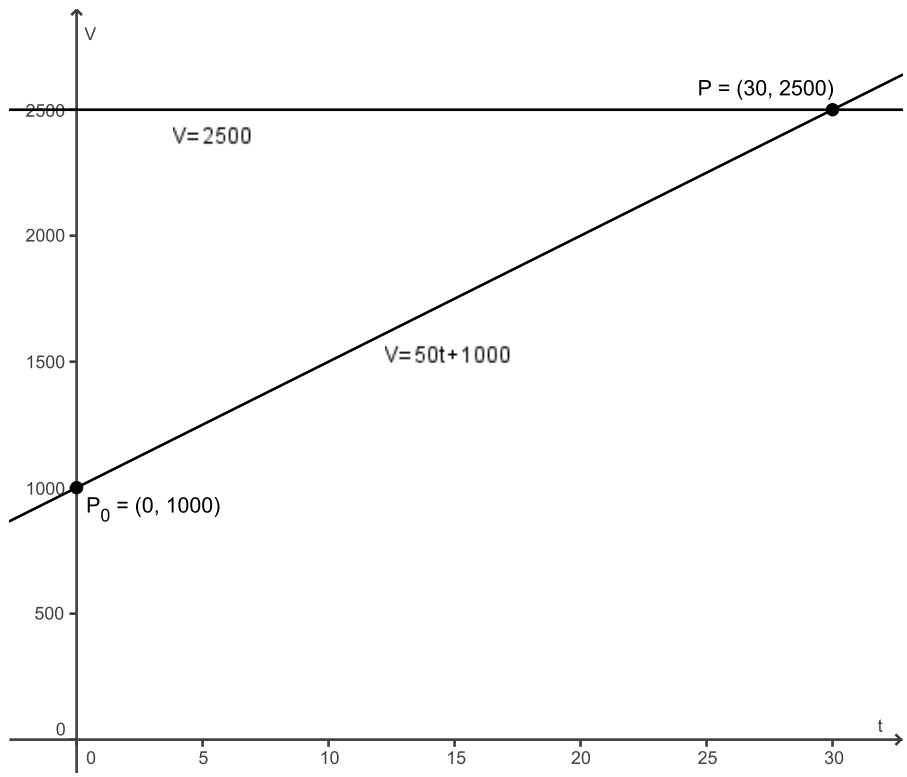

FIGURE 2.12. Solving the equation $50t + 1000 = 2500$ graphically

Again, there is a nice graphical solution. We can describe the water volume in terms of the time for the two different rates as follows:

$$V = 50t + 1000 \qquad \text{for the rate of } 50\ \ell/\text{min} \tag{2.25}$$

$$V = 30t + 1600 \qquad \text{for the rate of } 30\ \ell/\text{min} \tag{2.26}$$

The problem now asked for the time at which the volume is the same, i.e., to determine the t-coordinate of the point $P = (30, 2500)$ where the two lines described by the equations (2.25) and (2.26) intersect. (See Figure 2.13.)

A *general linear equation* has the form

$$m_0 x + b_0 = m_1 x + b_1 \tag{2.27}$$

and can be viewed graphically as finding the x-coordinate of the point at which the lines described by

$$y = m_0 x + b_0 \text{ and } y = m_1 x + b_1 \tag{2.28}$$

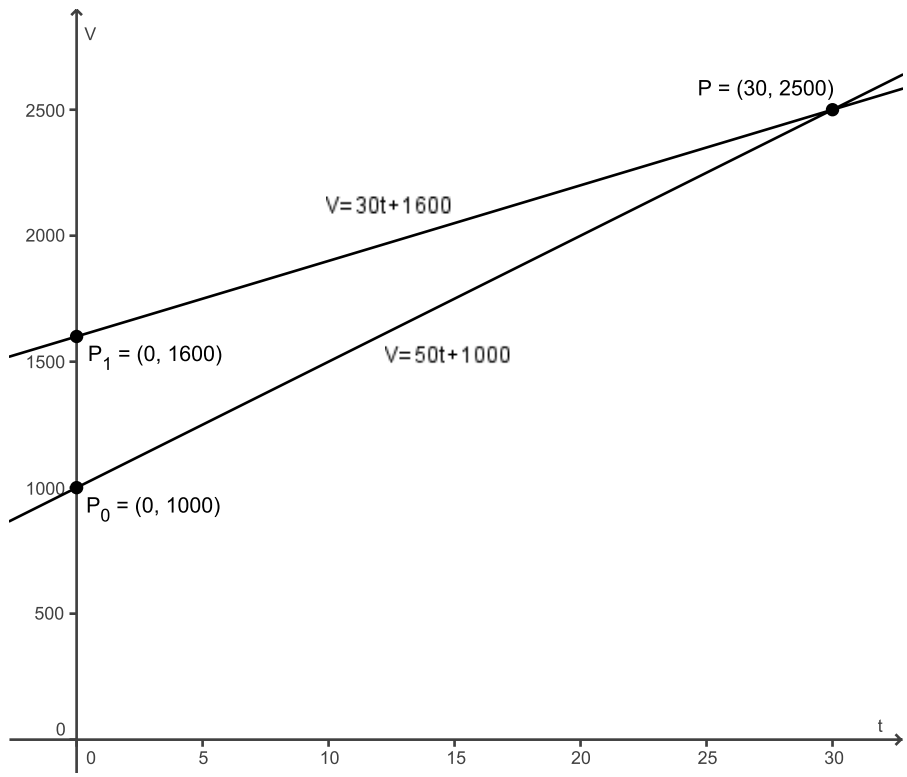

FIGURE 2.13. Solving the equation $50t + 1000 = 30t + 1600$ graphically

intersect (as, e.g., in Figure 2.12, where one line is horizontal, or Figure 2.13, where neither line is horizontal).

Algebraically speaking, we first subtract (up to two times) to "get all the x on one side" and "all the constants on the other side". In order to see what might happen, it's easier to view this problem graphically as finding the x-coordinate of the point where the two non-vertical lines described by the equations in (2.28) intersect. There are now three cases for the two lines described by the equations in (2.28):

Case 1: The lines intersect in a single point P, say: Then the two lines must have different slope, i.e., $m_0 \neq m_1$. In that case, there is exactly one solution to the equation (2.27), namely the x-coordinate

of P. Algebraically, we can find this value in three steps:

$$\begin{aligned}
(m_0 - m_1)\,x + b_0 &= b_1 && \text{(by subtracting } m_1 x\text{)}\\
(m_0 - m_1)\,x &= b_1 - b_0 && \text{(by subtracting } b_0\text{)}\\
x &= \frac{b_1 - b_0}{m_0 - m_1} && \text{(by dividing by } m_0 - m_1\text{)}
\end{aligned}$$

Here, the last step is possible since $m_0 \neq m_1$ and thus $m_0 - m_1 \neq 0$. A real-life example of such a situation would be the following:

> "Two trains leave the same city, each heading in the same direction at a constant speed of 80km/h and 120km/h, respectively. At noon, the first train is at a distance of 140km from the city, and the second train at a distance of 100km. At what time will the second train pass the first?

It is easy to see that solving this problem amounts to solving the equation $80t + 140 = 120t + 100$, and, after some algebra, one obtains the solution $t = 1$ (where the time t is measured in hours since noon). This solution is shown in Figure 2.14 as the point P at the intersection of the lines ℓ_1 and ℓ_0.

Case 2: The lines are parallel and have no point in common: Then the two lines must have the same slope, i.e., $m_0 = m_1$, but different y-intercept, i.e., $b_0 \neq b_1$. In that case, there is no solution to the equation (2.27). Algebraically, we can see this in three steps:

$$\begin{aligned}
(m_0 - m_1)\,x + b_0 &= b_1 && \text{(by subtracting } m_1 x\text{)}\\
(m_0 - m_1)\,x &= b_1 - b_0 && \text{(by subtracting } b_0\text{)}\\
0 &= b_1 - b_0 && \text{(since } m_0 = m_1\text{)}
\end{aligned}$$

The last equation is always false and thus has no solution x. A similar real-life example of such a situation would be the following:

> "Two trains leave the same city, both heading in the same direction at a constant speed of 120km/h. At noon, the first train is at a distance of 160km from the city, and the second train at a distance of 100km. At what time will the second train pass the first?

It is easy to see that solving this problem amounts to solving the equation $120t + 160 = 120t + 100$, which has no solution: The second train will never pass the first. This situation is also shown in Figure 2.14: The two lines ℓ_2 and ℓ_0 do not intersect.

Case 3: The lines are parallel and in fact coincide: Then the two lines must have the same slope and the same y-intercept, i.e., $m_0 =$

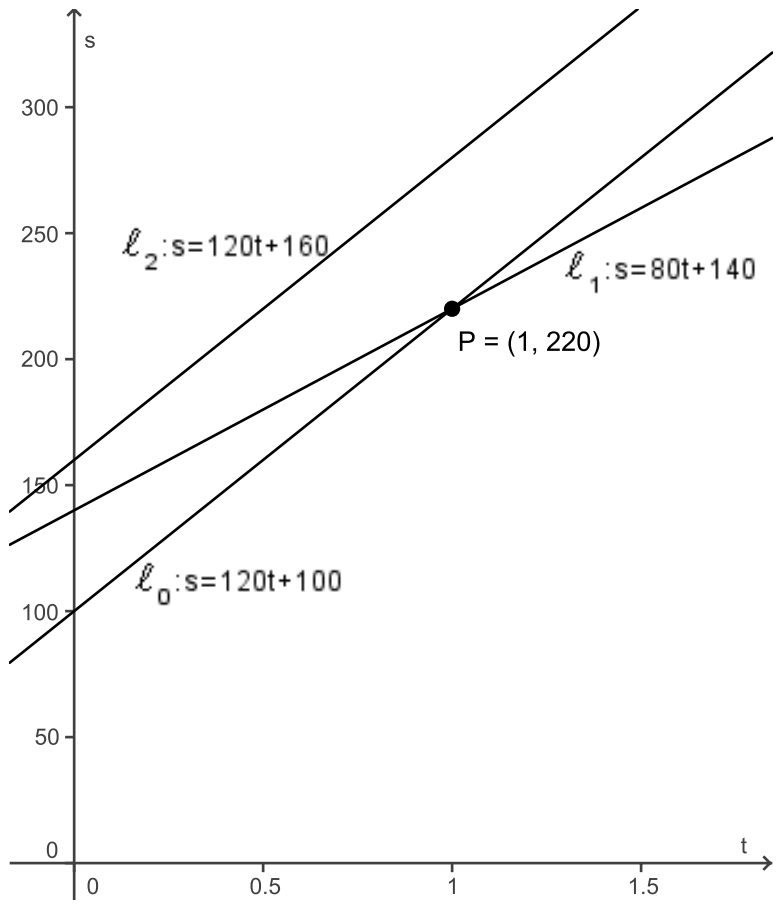

FIGURE 2.14. Solving the three trains problems graphically

m_1 and $b_0 = b_1$. In that case, every value of x is a solution to the equation (2.27). Algebraically, we can see this again in three steps:

$$\begin{aligned} (m_0 - m_1)\,x + b_0 &= b_1 && \text{(by subtracting } m_1 x\text{)} \\ (m_0 - m_1)\,x &= b_1 - b_0 && \text{(by subtracting } b_0\text{)} \\ 0 &= 0 && \text{(since } m_0 = m_1 \text{ and } b_0 = b_1\text{)} \end{aligned}$$

The last equation is always true, and so any value of x is a solution. A similar (somewhat non-sensical) real-life example of such a situation would be the following:

> "Two trains leave the same city, both heading in the same direction at a constant speed of 120km/h. At

> noon, both trains are at a distance of 100km from the city. At what time will one train pass the other?

It is easy to see that solving this problem amounts to solving the equation $120t + 100 = 120t + 100$, and that the trains will always be at the same distance from the city. This situation is shown in Figure 2.14: The travel of both trains is described by the same line ℓ_0.

2.6. Solving Simultaneous Linear Equations in Two or More Variables

Let's start again with an example:

> "I'm thinking of two numbers. The larger of the two numbers is less than twice the smaller number by 3. The sum of the two numbers is 18. What are my numbers?"

Now it is, of course, possible to solve this problem using only what we already know, namely, linear equations in one variable: If I call the smaller number x, then the larger number can be expressed as $2x - 3$ (since I'm told that "the larger of the two numbers is less than twice the smaller number by 3"). The other piece of information then tells us about the sum of the two numbers, namely, that

$$x + (2x - 3) = 18$$

which can be rewritten as $3x - 3 = 18$ and solved to $x = 7$ as discussed in the previous section. But it may, and will, not always be so easy to express one unknown quantity in terms of another; so an alternative and more general approach is to call the two numbers x and y (and let's say x is the smaller of the two). Now we can translate the information from the problem into two equations

$$\begin{aligned} y &= 2x - 3 \\ x + y &= 18 \end{aligned} \tag{2.29}$$

We can rewrite these equations into two line equations in general form as

$$\begin{aligned} -2x + y + 3 &= 0 \\ x + y - 18 &= 0 \end{aligned} \tag{2.30}$$

Solving these equations now means that we have to find (all the) values for the unknowns x and y which make both equations true.

There are a number of different methods to solve such *simultaneous linear equations* as we'll call them, i.e., several equations mentioning several unknowns (but where none of the unknowns is raised to any

power, or occurs in a denominator or in a product with another unknown) so that each equation is equivalent to a line equation in general form.

Geometrically, solving two simultaneous linear equations in two unknowns means that we need to find the point(s) of intersection of two lines in the plane. (If the lines are not parallel, there will be exactly one point of intersection.) Figure 2.15 shows the two lines from our example problem and the point of intersection $P = (7, 11)$.

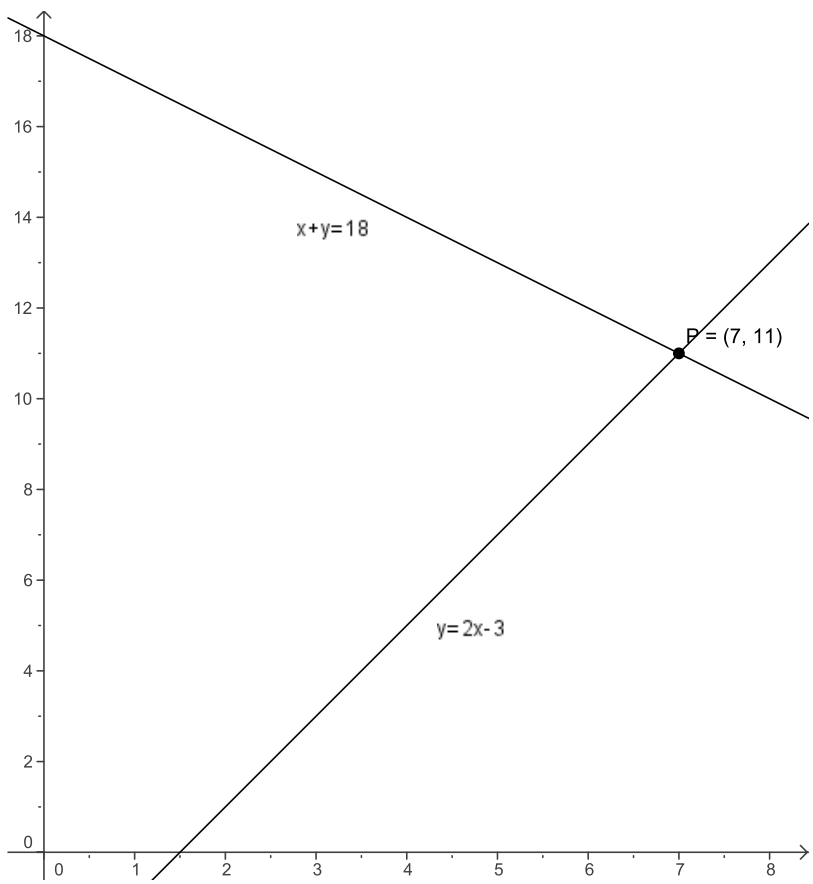

FIGURE 2.15. Solving the two equations $2x + y + 3 = 0$ and $x + y - 18 = 0$ graphically

Algebraically, there are a number of different ways to solve simultaneous linear equations, and experience will tell you and your students which method is more efficient in a particular situation. E.g., starting from (2.29), it would be easiest to use the *substitution method*: We already know that $y = 2x - 3$, so we could simply replace y in the other equation $x + y = 18$ by $2x - 3$ and obtain the single equation

$x + (2x - 3) = 18$ in a single unknown, which can then be solved as in the previous section. (This is essentially how we solved this problem at first at the beginning of the current section.)

We could also have started from (2.30), say, and found out from the equation $x + y = 18$ that $x = 18 - y$ (by subtracting y on both sides); now we can replace x by $18 - y$ in the other equation $-2x + y + 3 = 0$ to arrive at the equation $-2(18 - y) + y + 3 = 0$ in a single variable, which can then be simplified to $3y - 33 = 0$ and solved for y as in the previous section (namely, to $y = 11$). Clearly, the second way of using the substitution method is somewhat more cumbersome here, but the guiding principle is always the same: Rewrite one of the equations so that one variable occurs alone on one side and not at all on the other side, and then substitute it in the other equation to arrive at an equation in a single variable. Once we have found the value of one unknown, it's easy to find the value for the other unknown. E.g., in the substitution of the previous paragraph, we obtain $x = 7$ and then compute $y = 2 \cdot 7 - 3 = 11$. In the substitution of the current paragraph, we obtain $y = 11$ and then compute $x = 18 - 11 = 7$.

Starting from (2.30), there is a very different way to solve the equations called the *elimination method*. The idea is that the equation $x + y - 18 = 0$ is equivalent to $2x + 2y - 36 = 0$ (by multiplying both sides of the equation by 2). Now we can *add* this new equation $2x + 2y - 36 = 0$ and the other equation $-2x + y + 3 = 0$ and arrive at $(2x + 2y - 36) + (-2x + y + 3) = 0 + 0$. After simplifying to $3y - 33 = 0$, we see that we have "eliminated" one unknown, namely, x, and so have an equation we can again solve by the method of the previous section.

Alternatively, we could have *subtracted* the first equation $-2x + y + 3 = 0$ in (2.30) from the second equation $x + y - 18 = 0$ to arrive at $(x + y - 18) - (-2x + y + 3) = 0 - 0$, and after simplifying to $3x - 21 = 0$, we have again "eliminated" one unknown, namely, y this time, and so have an equation we can again solve by the method of the previous section.

In either way of eliminating one unknown, we can solve the resulting equation in one unknown and then replace the solution for this unknown in one of the original equations to now solve for the other unknown. E.g., in the last paragraph, once we have solved the new equation $3x - 21$ to $x = 7$, we can solve for y by solving $7 + y - 18 = 0$ and arrive at $y = 11$.

Note that the elimination method introduces a new way to rewrite equations. In Proposition 2.6, we had discussed rewriting a single equation by adding or subtracting the same expression to or from both sides,

or by multiplying or dividing both sides by the same nonzero expression on both sides. The elimination method requires a new operation on *two simultaneous* equations:

PROPOSITION 2.8. *(1) Any solutions to two simultaneous equations $A = B$ and $C = D$ (where A, B, C, and D are algebraic expressions) remain solutions if we add or subtract one equation to (or from) the other, i.e., the simultaneous solutions of $A = B$ and $C = D$ are still simultaneous solutions of the simultaneous equations $A + C = B + D$ and $C = D$, and also simultaneous solutions of the simultaneous equations $A - C = B - D$ and $C = D$. (Similarly, for later reference, if C and D are nonzero expressions, then any solutions to two simultaneous equations $A = B$ and $C = D$ (where A, B, C, and D are algebraic expressions) remain solutions if we multiply or divide one equation by the other, i.e., the simultaneous solutions of $A = B$ and $C = D$ are still simultaneous solutions of the simultaneous equations $AC = BD$ and $C = D$, and also simultaneous solutions of the simultaneous equations $\frac{A}{C} = \frac{B}{D}$ and $C = D$.)*

(2) Conversely, any simultaneous solutions of the equations $A + C = B + D$ and $C = D$ (or $A - C = B - D$ and $C = D$, respectively) are also simultaneous solutions of the original equations $A = B$ and $C = D$. (Similarly, again for later reference, if C and D are nonzero expressions, then any simultaneous solutions of the equations $AC = BD$ and $C = D$ (or $\frac{A}{C} = \frac{B}{D}$ and $C = D$, respectively) are also simultaneous solutions of the original equations $A = B$ and $C = D$.)

The proof of the part (1) of Proposition 2.8 is simple: If we replace the variables in A, B, C, and D by any fixed numbers, then the truth of the equations $A = B$ and $C = D$ with the variables replaced by these fixed numbers implies the truth of the equations $A+C = B+D$, $A-C = B-D$, $AC = BD$ and $\frac{A}{C} = \frac{B}{D}$. For part (2) of Proposition 2.8, we have to work just a little harder. E.g., for the case $A+C = B+D$ and $C = D$, we use the fact that we can "reverse" the operation of "adding equations" by "subtracting": From $A + C = B + D$ and $C = D$, we obtain $(A + C) - C = (B + D) - D$, which simplifies to the desired equation $A = B$. Similarly, for the case $AC = BD$ and $C = D$, we use the fact that we can "reverse" the operation of "multiplying equations" by "dividing": From $AC = BD$ and $C = D$, we obtain $\frac{AC}{C} = \frac{BD}{D}$, which simplifies to the desired equation $A = B$, given that C and D are nonzero expressions.

Let's now return to our geometric intuition for solving two simultaneous linear equations in two unknowns again to see what else can happen. We had said that, geometrically, solving two simultaneous linear equations in two unknowns amounts to finding the point(s) of intersection of the two lines described by the two equations. This is easy by either of the two algebraic methods (substitution method or elimination method) described above *as long as* the two lines are not parallel and therefore intersect in exactly one point. Let's vary our original problem a bit as follows:

> "I'm thinking of two numbers. The sum of the two numbers is 12. The sum of twice each of the two numbers exceeds 24 by 3. What are my numbers?"

We can quickly translate this problem into the following two simultaneous equations

$$\begin{aligned} x + y &= 12 \\ 2x + 2y - 3 &= 24 \end{aligned} \tag{2.31}$$

If we tried to solve (2.31) by the substitution method, we could rewrite the first equation as $x = 12 - y$ and thus the second equation as $2(12 - y) + 2y - 3 = 24$. But this last equation simplifies to $24 - 3 = 24$, which is always false. Since this last equation has no simultaneous solutions x and y, neither do the original equations (2.31), by Proposition 2.8 (2).

Or we could have tried to solve (2.31) by the elimination method, multiplying the first equation by 2 to obtain $2x + 2y = 24$, and then subtracting that from the second equation to arrive at $(2x+2y)-(2x+2y-3) = 24 - 24$, which simplifies to $3 = 0$, again an equation which is always false. And this implies that the original equations (2.31) do not have any simultaneous solutions, again by Proposition 2.8 (2).

So what happened? If we rewrite (2.31) as line equations in slope-intercept form

$$\begin{aligned} y &= -x + 12 \\ y &= -x + 13.5 \end{aligned} \tag{2.32}$$

we quickly see that these lines are parallel and do not intersect; so it is also geometrically clear that they cannot have any simultaneous solutions.

Finally, it is possible that the two lines from two simultaneous linear equations coincide, in which case there are infinitely many simultaneous solutions. Here is a somewhat non-sensical example:

> "I'm thinking of two numbers. The sum of the two numbers is 12. The sum of twice each of the two numbers equals 24. What are my numbers?"

This problem can be expressed by the two simultaneous equations

$$\begin{aligned} x + y &= 12 \\ 2x + 2y &= 24 \end{aligned} \tag{2.33}$$

But these equations represent the same line, with the equation $y = -x + 12$ in slope-intercept form, so any solution x is possible as long as y is then chosen so as to satisfy $y = -x + 12$.

Finally, the above carries over to simultaneous linear equations in more than two unknowns. Here is an example:

> "Richard spends \$35 on two books, a CD, and a DVD. Susan can only afford to pay \$25 for a CD and a DVD. And Tom pays just \$20 for a book and a DVD. Assuming that each book costs the same, each CD costs the same, and each DVD costs the same, what is the price of each item?"

Let's denote the price of a book by b, the price of a CD by c, and the price of a DVD by d. Then we arrive at the equations

$$\begin{aligned} 2b + c + d &= 35 \\ c + d &= 25 \\ b + d &= 20 \end{aligned} \tag{2.34}$$

representing in turn the total cost of Richard's, Susan's, and Tom's purchases, respectively.

It is "possible" to visualize a solution to this problem in three-dimensional coordinate space: Each linear equation then represents a *plane* in three-dimensional space, and a solution is given by any point at which all three planes intersect. Obviously, this is rather hard to draw (or even visualize), so school mathematics generally approaches this kind of problem only algebraically by one of our two methods, the substitution method or the elimination method, or a combination of the two. Here are two (of several) possible ways to solve (2.34) algebraically:

(1) The second and third equation of (2.34) allow us to express both c and b in terms of d by subtracting d on both sides of each equation, arriving at

$$\begin{aligned} c &= 25 - d \\ b &= 20 - d \end{aligned}$$

Now we can replace c and b in the first equation of (2.34) and obtain the following equation in only one unknown:

$$2(20-d)+(25-d)+d=35$$

This equation simplifies to

$$65-2d=35$$

and gives us the solution $d=15$ as in the previous section. Now we can also compute

$$c=25-d=25-15=10$$
$$b=20-d=20-15=5$$

(2) We can start by subtracting the second equation of (2.34) from the first and arrive at the equation

$$(2b+c+d)-(c+d)=35-25$$

which simplifies to the following equation in only one unknown:

$$2b=10$$

This gives us the solution $b=5$. Now we can use the third equation of (2.34) to compute

$$d=20-b=20-5=15$$

and then use the second equation of (2.34) to compute

$$c=25-d=25-15=10$$

In either case, we see that a book costs \$5, a CD \$10, and a DVD \$15.

CHAPTER 3

Order and Linear Inequalities

3.1. Ordering the Numbers

All the number systems we will consider here (except for the complex numbers, which we will touch upon only briefly in subsection 5.3.2) carry a natural "order", which formalizes the very intuitive idea of "more" and "less". This concept is introduced very early for the whole numbers: If Mary has seven apples and John has five, then Mary has more apples than John. Using like denominators, one can carry over this concept to fractions; and using place value, to positive decimals.

On the other hand, the concept of "more" and "less" is somewhat counterintuitive for negative numbers and doesn't always conform with everyday language: If Susan owes \$5 and Bob owes \$8, then Bob "owes more" than Susan; but if one considers their "savings" and allows this to be negative in the case of debt, then Susan has "more" savings than Bob. A better model for negative numbers and their order is temperature: $-20°$ is less than $-10°$ and also less than $+10°$ (even though 20 is "bigger" than 10). Another intuitive way to think about the order of all numbers (positive, negative, or 0) is on the number line: Being less than a number means being to the left of the point on the number line representing that number.

Such concepts are expressed in mathematics by inequalities, which is somewhat of a misnomer: We express that two numbers, say, 2 and 3, are *unequal* by writing $2 \neq 3$, which is literally an "inequality"; but one *usually* understands an *inequality* to be a statement involving one of the symbols $<$, $>$, $\leq$, or $\geq$. We will start by "formally" using only $<$, and we will consider $a > b$ to be merely an *abbreviation* for $b < a$.

Here are some basic properties of the ordering $<$ for all numbers a, b, and c:

(irreflexivity)	$a < a$ is false
(transitivity)	$a < b$ and $b < c$ implies $a < c$
(comparability)	$a < b$ or $a = b$ or $a > b$

So the first property (irreflexivity) states that no number is less than "itself" (thus the name of the property: "reflexive" means "referring

to itself"). The second property (transitivity) states that from two facts one can conclude a third: If I know that b is greater than a but less than c, then certainly a is less than c. The third property (comparability) states that whenever two numbers a and b are *not* equal, then one must be less than the other.

What is a bit confusing is that there is an alternative way to state properties of the ordering of the numbers, using the symbol $\leq$. (Again, $a \geq b$ is simply an *abbreviation* for $b \leq a$, so "formally", we don't use $\geq$ at all!) Now there are two ways to deal with the symbol $\leq$:

On the one hand, we could simply consider $a \leq b$ as an abbreviation by letting

$$a \leq b \qquad \text{abbreviate} \qquad a < b \text{ or } a = b \tag{3.1}$$

and so the symbol $\leq$ gives no additional meaning which we couldn't express before.

On the other hand, we could *replace* the symbol $<$ by the symbol $\leq$ and *only* deal with the latter. Then the basic properties of the ordering under $\leq$ become the following, for all numbers a, b, and c:

(reflexivity)	$a \leq a$
(transitivity)	$a \leq b$ and $b \leq c$ implies $a \leq c$
(antisymmetry)	$a \leq b$ and $b \leq a$ implies $a = b$
(comparability)	$a \leq b$ or $a \geq b$

So now the first property (reflexivity) states that every number is less than or equal to "itself" (thus again the name of the property). The second property (transitivity) states that from two facts one can conclude a third: If I know that b is greater than or equal to a but less than or equal to c, then certainly a is less than or equal to c. The third property (antisymmetry) states that if a number is both less than or equal to, and greater than or equal to, another number, then the two numbers are indeed equal. The last property (comparability) states that for any two numbers, one is less than or equal the other, or vice versa.

Now, assuming that we have only introduced the symbol $\leq$ so far (but *not* $<$), we can think of $<$ as an abbreviation by letting

$$a < b \qquad \text{abbreviate} \qquad a \leq b \text{ and } a \neq b \tag{3.2}$$

So, in summary, it doesn't really matter which of the two symbols, $<$ or $\leq$, we introduce first; we can simply define the other from it. And in fact the properties of $\leq$ can be proved from the properties of $<$ and vice versa:

Exercise 3.1. (1) Show that assuming the three properties of $<$ and the definition of $\leq$ by (3.1), we can prove the four properties of $\leq$.

(2) Conversely, show that assuming the four properties of $\leq$ and the definition of $<$ by (3.2), we can prove the three properties of $<$.

Two more remarks are in order: In light of the comparability properties, we don't need any symbols like $\not<$ or $\not\leq$, since we can rewrite, e.g., $a \not< b$ as $a \geq b$ (i.e., as $b \leq a$), and $a \not\leq b$ as $a > b$ (i.e., as $b < a$).

Finally, there are many different ways to express order in English, using words like "exceed", "not", etc.:

Exercise 3.2. For each of the statements $a < b$, $a \leq b$, $a > b$, and $a \geq b$, find at least *four different* ways to express these in the English language, for quantities a and b.

3.2. Order and the Arithmetical Operations

So far, we have considered the ordering of the numbers and the arithmetical operations separately. But they are closely connected indeed, as we will now explore in detail. These connections will be essential in "solving inequalities" later on.

First of all, recall the following trivial properties linking *equality* of numbers and the arithmetical operations. (These were crucial in establishing Propositions 2.6 and 2.8!)

$$a = b \text{ implies } a + c = b + c \tag{3.3}$$

$$a = b \text{ implies } a - c = b - c \tag{3.4}$$

$$c \neq 0 \text{ and } a = b \text{ implies } ac = bc \tag{3.5}$$

$$c \neq 0 \text{ and } a = b \text{ implies } \frac{a}{c} = \frac{b}{c} \tag{3.6}$$

In fact, not only do these properties hold, but it's also not hard to see that their converses hold; i.e., the right-hand side of each statement implies the left-hand side (always assuming that $c \neq 0$ in (3.5) and (3.6)). This is since we can "undo" the addition in (3.3) using (3.4); the subtraction in (3.4) using (3.3); the multiplication in (3.5) using (3.6); and the division in (3.6) using (3.5).

There are now the following similar properties linking the order $<$ of the numbers and the arithmetical operations:

$$a < b \text{ implies } a + c < b + c \tag{3.7}$$

$$a < b \text{ implies } a - c < b - c \tag{3.8}$$

$$c > 0 \text{ and } a < b \text{ implies } ac < bc \tag{3.9}$$

$$c > 0 \text{ and } a < b \text{ implies } \frac{a}{c} < \frac{b}{c} \tag{3.10}$$

Why do properties (3.7)–(3.10) hold? This is easy and intuitive to see for properties (3.7)–(3.8): Adding or subtracting c to both sides of an inequality does not change its truth. Similarly, for properties (3.9)–(3.10), multiplying or dividing both sides of an inequality by a positive number c does not change its truth.

There are also more visual proofs: The implications (3.7) and (3.8) hold since we can think of them as amounting to "shifting" a and b on the number line to the right or left by an equal distance, given by c. The implications (3.9) and (3.10) hold since we can think of them as amounting to "scaling" a and b on the number line by the positive factor or divisor c.

Again, we have the converses of these properties (the "right-to-left" direction of these implications), "undoing" the operations as for (3.3)–(3.6), still under the assumption that $c > 0$ in (3.9) and (3.10).

We can also state similar properties and their converses using $\leq$, except that we still assume $c > 0$ in (3.13) and (3.14):

$$a \leq b \text{ implies } a + c \leq b + c \tag{3.11}$$

$$a \leq b \text{ implies } a - c \leq b - c \tag{3.12}$$

$$c > 0 \text{ and } a \leq b \text{ implies } ac \leq bc \tag{3.13}$$

$$c > 0 \text{ and } a \leq b \text{ implies } \frac{a}{c} \leq \frac{b}{c} \tag{3.14}$$

And again we have the converses of these properties (the "right-to-left" direction of these implications), "undoing" the operations as for (3.3)–(3.6), still under the assumption that $c > 0$ in (3.13) and (3.14).

But this leads to an obvious question, and one which is at the root of many errors when solving inequalities: Why do we need the restriction that $c > 0$ in properties (3.9)–(3.10) and (3.13)–(3.14)? We start with the following easy but crucial observations:

$$a < b \text{ implies } -a > -b \tag{3.15}$$

$$a \leq b \text{ implies } -a \geq -b \tag{3.16}$$

Note that (3.15) and (3.16) hold not only for positive numbers a and b, but for 0 and negative numbers as well!

In order to see why (3.15) holds, we distinguish three cases:

Case $0 \leq a < b$*:* Then b lies to the right of a on the number line, and both are ≥ 0. Since taking the negative of a number means to reflect the points on the number line about 0, $-b$ will now lie to the left of a on the number line, and both are ≤ 0. (See Figure 3.1.)

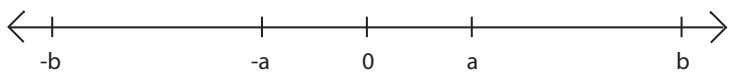

FIGURE 3.1. The proof of (3.15) when $0 \leq a < b$

Case $a < b \leq 0$*:* Then b lies to the right of a on the number line, and both are ≤ 0. Since taking the negative of a number means to reflect the points on the number line about 0, $-b$ will now lie to the left of $-a$ on the number line, and both are ≥ 0. (See Figure 3.2.)

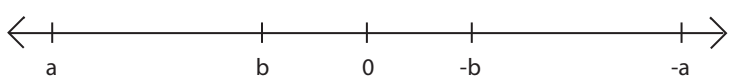

FIGURE 3.2. The proof of (3.15) when $a < b \leq 0$

Case $a < 0 < b$*:* Then b lies to the right of 0, and a to the left of 0, on the number line. Since taking the negative of a number means to reflect the points on the number line about 0, $-b$ will now lie to the left of 0, and $-a$ to the right of 0, on the number line. (See Figure 3.3.)

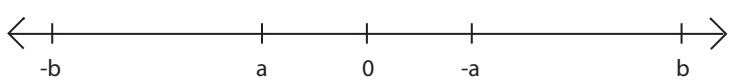

FIGURE 3.3. The proof of (3.15) when $a < 0 < b$

In each case, we have established $-b < -a$ as required for (3.15). The proof for (3.16) is similar.

Using (3.15), we can now prove versions of (3.9) and (3.10) as well as (3.13) and (3.14) when $c < 0$:

$$c < 0 \text{ and } a < b \text{ implies } ac > bc \tag{3.17}$$

$$c < 0 \text{ and } a < b \text{ implies } \frac{a}{c} > \frac{b}{c} \tag{3.18}$$

$$c < 0 \text{ and } a \leq b \text{ implies } ac \geq bc \tag{3.19}$$

$$c < 0 \text{ and } a \leq b \text{ implies } \frac{a}{c} \geq \frac{b}{c} \tag{3.20}$$

Each of (3.17)–(3.20) follows from the corresponding version when $c > 0$ by applying (3.15) or (3.16). E.g., (3.17) is shown as follows, assuming $c < 0$:

$$\begin{aligned} a < b \quad & \text{implies} \quad a(-c) < b(-c) \\ & \text{implies} \quad a(-1)(-c) > b(-1)(-c) \\ & \text{implies} \quad ac > bc \end{aligned} \tag{3.21}$$

The first implication follows by (3.9) (using that $-c > 0$), and the second by (3.15) (using that $-(a(-c)) = a(-1)(-c)$), while the third just evaluates $(-1)(-c) = c$.

A more visual proof of (3.17), closely following (3.21), goes as follows: Multiplying an inequality $a < b$ by a negative number c amounts to first scaling a and b on the number line by the *positive* number $-c$ (as in the first line of (3.21)), followed by a reflection of the number line about the number 0 (as in the second line of (3.21)).

Again, we have the converses of (3.17)–(3.20) (i.e., the "right-to-left" implications also hold, always assuming that $c < 0$) since we can "undo" the multiplication or division by c by dividing or multiplying by $\frac{1}{c}$. This uses that $c < 0$ implies $\frac{1}{c} < 0$, which follows by dividing both sides of $c < 0$ by $\frac{1}{c}$ twice.

3.3. Linear Inequalities in One Unknown

Equipped with the properties from the last section, it is now easy to solve *linear inequalities* in one unknown, i.e., inequalities where both sides are linear expressions. Let's start with an example:

> Little Red Riding Hood falls asleep in the forest, 5km away from her grandmother's house. The next morning, she walks toward her grandmother's house at a constant speed of 3km/h for a while. If she is now at most 2km away from her grandmother's house, how long will she have walked?

Let's denote the time (in hours, from when she starts walking) by t and the distance from her grandmother's house (in kilometers) by s. Then we have the following equation for time and distance:

$$s = -3t + 5 \tag{3.22}$$

The problem tells us that this quantity is at most 2 at the end of the walk, leading to the inequality

$$-3t + 5 \leq 2$$

Now we can first subtract 5 from both sides (using (3.12)) to obtain

$$-3t \leq -3$$

and then divide by -3 (using (3.20)) to arrive at the solution

$$t \geq 1 \tag{3.23}$$

Thus she will have walked at least one hour.

Note that solving linear inequalities is thus very similar to solving linear equations, the one difference being that whenever we multiply or divide by a negative number, we have to invert the inequality sign. This is reflected in the following proposition, which closely follows Proposition 2.6:

PROPOSITION 3.3. *(1) Any solution to an inequality $A < B$ (where A and B are algebraic expressions) remains a solution if we add or subtract the same algebraic expression on both sides of the inequality; and vice versa, any solution to the inequality* after *adding or subtracting the same algebraic expression is also a solution of the original inequality $A < B$.*

(2) Any solution to an inequality $A < B$ (where A and B are algebraic expressions) remains a solution if we multiply or divide both sides of the inequality by the same positive *algebraic expression; and vice versa, any solution to the inequality* after *multiplying or dividing by the same positive algebraic expression is also a solution of the original inequality $A < B$. (Here, an algebraic expression is* positive *if it is positive no matter what numbers replace the letters in the expression.)*

(3) Any solution to an inequality $A < B$ (where A and B are algebraic expressions) is a solution of the inequality $AC > BC$ (or $\frac{A}{C} > \frac{B}{C}$) obtained by multiplying (or dividing, respectively) both sides of the inequality by the same negative *algebraic expression and switching the inequality sign; and vice versa, any solution to the inequality $AC > BC$ (or $\frac{A}{C} > \frac{B}{C}$)* after *multiplying (or dividing, respectively) by the same negative algebraic*

expression and switching the inequality sign is also a solution of the original inequality $A < B$. (Here, an algebraic expression is negative *if it is negative no matter what numbers replace the letters in the expression.)*

These results also hold with $<$ and $>$ switched; with $<$ and $>$ replaced by $\leq$ and $\geq$, respectively; and with $<$ and $>$ replaced by $\geq$ and $\leq$, respectively.

The proof of Proposition 3.3 is basically the same as for Proposition 2.6: Adding or subtracting an algebraic expression on both sides of an inequality, or multiplying and dividing both sides of an inequality by a nonzero algebraic expression, does not affect the solutions by the results of Section 3.2 and the fact that all these operations on inequalities can be "undone", except for the caveat that multiplication or division by a negative number flips the inequality symbol.

Here is now a slightly more complicated example:

> Little Red Riding Hood falls asleep in the forest, 5km away from her grandmother's house. The next morning, she starts walking toward her grandmother's house at a constant speed of 3km/h. When she starts walking, the big bad wolf starts along the same trail at a constant speed of 6km/h from a distance of 8km from the grandmother's house. How long will Little Red Riding Hood be safe from the wolf?

Let's denote the time (in hours, from when she starts walking) by t and the distance from her grandmother's house (in kilometers) by s_R and s_W for Little Red Riding Hood and the big bad wolf, respectively. Then we have the following equation relating time and distance for Little Red Riding Hood:

$$s_R = -3t + 5$$

And for the big bad wolf, the equation relating time and distance reads:

$$s_W = -6t + 8$$

The problem tells us that we want to find the times t when

$$s_R < s_W$$

since Little Red Riding Hood is only safe from the wolf as long as she is closer to her grandmother's house than the wolf. This leads to the inequality

$$-3t + 5 < -6t + 8 \tag{3.24}$$

Now we can first subtract 5 from both sides (using (3.8)) to obtain

$$-3t < -6t + 3$$

then add $6t$ to both sides (using (3.7)) to obtain

$$3t < 3$$

and finally divide by 3 (using (3.10)) to arrive at the solution

$$t < 1 \tag{3.25}$$

Thus she will be safe at any time less than one hour.

There are two other nice ways (in addition to (3.23) and (3.25)) to represent the solutions to inequalities like (3.22) and (3.24):

- We can represent the solutions as intervals using the notation from subsection 1.5.1: The solutions to (3.22) are all $t \geq 1$, and so the *solution set* can be written as $[1, \infty)$. Similarly, the solutions to (3.24) are all $t < 1$, and so the *solution set* can be written as $(-\infty, 1)$.
- We can represent the solutions on the number line, by letting the points corresponding to solutions be denoted by the thicker part of the number line, and by using solid dots for interval endpoints which are a solution, and circles for interval endpoints which are not a solution. The solutions to (3.22) and (3.24) on the number line are now pictured in Figures 3.4 and 3.5, respectively.

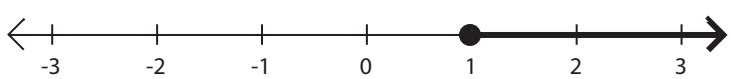

FIGURE 3.4. The solution set of (3.22) on the number line

Here is one final example, illustrating how more than one inequality for a quantity works:

> Little Red Riding Hood falls asleep in the forest, 5km away from her grandmother's house. The next morning, she walks toward her grandmother's house at a constant speed of 3km/h for some time. How long

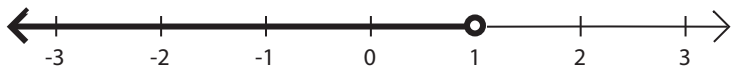

FIGURE 3.5. The solution set of (3.24) on the number line

> will she have walked when she is between 3km and 2km from her grandmother's house?

The equation relating time and distance for Little Red Riding Hood still reads

$$s = -3t + 5$$

where s is the distance from grandmother's house (in kilometers) and t the time from when she starts walking (in hours). The information from the problem now tells us that the distance is between 2 and 3, leading to the inequalities

$$2 \leq -3t + 5 \text{ and } 3 \geq -3t + 5 \tag{3.26}$$

which is usually written more compactly as

$$2 \leq -3t + 5 \leq 3 \tag{3.27}$$

Note that this compact notation only works since the inequalities in (3.26) are linked by the word "and"! In general, you can abbreviate the statement "$A \leq B$ and $B \leq C$" by "$A \leq B \leq C$"; and similarly for $A < B < C$, $A \geq B \geq C$, and $A > B > C$. More about this in a minute!

We can perform the same operations on the inequality (3.27) to solve it: First subtract 5 on all sides to obtain

$$-3 \leq -3t \leq -2$$

and then divide by -3 on all sides to arrive at the solution

$$1 \geq t \geq \frac{2}{3} \tag{3.28}$$

So Little Red Riding Hood walks between 40 minutes and one hour. Again, we can represent the solution set by an interval, namely, $[\frac{2}{3}, 1]$, or on the number line as in Figure 3.6.

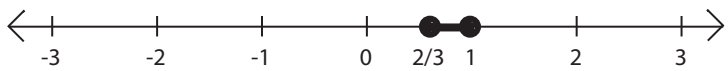

FIGURE 3.6. The solution set of (3.27) on the number line

Note that in this last example, the set of t which are *not* a solution can be described by the statement

$$2 > -3t + 5 \text{ or } 3 < -3t + 5$$

which is frequently (but incorrectly!) abbreviated as $3 < -3t + 5 < 2$. This last statement makes no sense at all since it would imply that $3 < 2$. The "real" problem is that one cannot write the set of "non"-solutions of (3.27) as a single interval; rather, it consists of two intervals, namely, $(-\infty, \frac{2}{3})$ and $(1, \infty)$. Using our set notation from subsection 1.5.1, we can also write the set of "non"-solutions of (3.27) as $(-\infty, \frac{2}{3}) \cup (1, \infty)$.

We will return to more complicated inequalities later on and then discuss possible solution sets of inequalities in more detail. For now, we turn to one common application of inequalities, estimation and approximation.

3.4. Estimation and Approximation

In a "purely mathematical" setting, the number 2 means exactly that: It's the number 2, no more and no less. And writing it differently, e.g., as 2.0 or 2.00, does not change that fact.

But in the "real world", we can't measure quantities *exactly*; we can only measure them *approximately*. And there often is a "convention" that 2 means 2 rounded to the nearest integer (i.e., 2 denotes a quantity x which satisfies the inequality $1.5 \leq x < 2.5$), whereas 2.0 means 2 rounded to the nearest tenth (i.e., 2.0 denotes a quantity x which satisfies the inequality $1.95 \leq x < 2.05$), and 2.00 means 2 rounded to the nearest hundredth (i.e., 2.00 denotes a quantity x which satisfies the inequality $1.995 \leq x < 2.005$).

A better and more precise way to write these would be to write $2 \pm .5$; $2 \pm .05$; and $2 \pm .005$, respectively. One advantage is that it allows us to write things like $2 \pm .1$ (i.e., 2 "rounded to the closest

fifth" or, more precisely, a quantity x which satisfies the inequality $1.9 \leq x < 2.1$). Another advantage is that when reading 120, we can't really be sure whether this means 120 "rounded to the closest integer" (i.e., a quantity x which satisfies the inequality $119.5 \leq x < 120.5$), or 120 "rounded to the closest ten" (i.e., a quantity x which satisfies the inequality $115 \leq x < 125$); whereas, reading it as $120 \pm .5$ or 120 ± 5, respectively, clarifies which of these two we mean.

There are now some quirks about estimation and approximation which are frequently swept under the rug in school mathematics but which are important to keep in mind for you as a teacher, and for your students to at least be aware of, especially in the age of ubiquitous pocket calculators which can spew out a number with eight significant digits regardless of how imprecise the input was!

We'll restrict ourselves here to just one example with two parts:

> Given a rectangle of length 8.4cm and width 5.2cm, what are (1) its perimeter, and (2) its area? Now suppose that length and width are only given to the nearest tenth, what are the possible values for (1) its perimeter and (2) its width?

Your first inclination may be to reach for a pocket calculator and to compute the perimeter as

$$p = 2 \cdot 8.4 + 2 \cdot 5.2 = 27.2$$

and the area as

$$A = 8.4 \cdot 5.2 = 43.68$$

so, purely mathematically, the perimeter is exactly 27.2cm and the area exactly 43.68cm^2. But in "real life", the length l measuring 8.4cm and the width w measuring 5.2cm means that

$$8.35 \leq l < 8.45 \text{ and } 5.15 \leq w < 5.25$$

For the perimeter, this implies that

$$27 = 2 \cdot 8.35 + 2 \cdot 5.15 \leq 2 \cdot l + 2 \cdot w < 2 \cdot 8.45 + 2 \cdot 5.25 = 27.4$$

and so $p = 27.2 \pm .2$ in our earlier notation. Similarly, for the area,

$$43.0025 = 8.35 \cdot 5.15 \leq l \cdot w < 8.45 \cdot 5.25 = 44.3625$$

and so the area can vary almost all the way between 43cm^2 and 44.4cm^2! (Using the $\pm$ notation here is very messy. And in general, it's quite hard to give the exact bounds.)

We therefore see that especially computing the area as (exactly) 43.68cm^2 is highly misleading! The general rule of thumb is that one

should never give more significant digits for the output than for the inputs: Under this rule of thumb, we would have computed the perimeter as 27cm and the area as 44cm^2 (rounding each to two significant digits), which is not quite accurate by our computations above (since the latter suggests that $26.5 \leq p < 27.5$ and $43.5 \leq A < 44.5$), but at least it is "in the right ballpark" for the accuracy of the answer.

CHAPTER 4

The Concept of a Function, and Functions Closely Related to Linear Functions

4.1. The Concept of a Function

In Chapter 2 on linear equations, we already saw examples of how one quantity (let's denote it by y) varies depending on another quantity (which we will call x for now). In particular, the equation for a (non-vertical) line can always be written in the form

$$y = mx + b \tag{4.1}$$

for some fixed parameters m and b. In such a situation, we call x the *independent variable*, and y the *dependent variable*, since we consider how "y varies as x varies".

First note that in the particularly simple case of (4.1) (and as long as $m \neq 0$), the choice of which variable to call dependent and which independent is somewhat arbitrary from a mathematical point of view since (4.1) can be rewritten as

$$x = \frac{y}{m} - \frac{b}{m}$$

However, the "real life" context might make one choice more "natural" than the other: Say, x might be the time elapsed since a train left the station at a constant speed, and y might be the distance of the train from the station: It would then be more natural to consider the distance as time varies, rather than the time as distance varies, since time "can't be made flexible" whereas the distance of the train can vary in a different pattern if the speed equals a different constant, or is no longer constant at all.

Putting this minor point aside, let's concentrate on the concept of how a quantity y varies depending on how another quantity x varies: Mathematics usually expresses such a situation by the concept of a *function*

$$y = f(x) \tag{4.2}$$

First of all, note that this notation may already cause some confusion: $f(x)$ does *not* mean that we multiply quantity f by quantity x,

but rather that, as a computer scientist might put it, we apply some "rule" f to the input x in order to determine the output y.

The set of possible values for the quantity x is called the *domain* of the function; it is determined either mathematically by when the equation $y = f(x)$ "makes sense" or "is defined", or by the "real life" context as to when it makes sense for the "real life" quantity x. The critical feature of a function is, however, the fact that for each argument x in the domain of the function, there is *exactly one* value for y such that $y = f(x)$ holds, as already suggested by the way we write this. But note that we could have written the relation between the quantities x and y differently, and it wouldn't have been so easy to check whether the relation defines a function: E.g., the relation

$$x - y^2 = 0 \tag{4.3}$$

can be expressed as

$$x = y^2$$

and so we can think of it as describing a function for the quantity x as y varies. (The fact that I "switched" the roles of x and y here is deliberate: The choice of what to call the variables is also completely arbitrary, and one should not always choose the same letter for the independent and the dependent variable, respectively!) On the other hand, (4.3) does *in general* not describe y as a function of x: E.g., setting $x = 1$, both $y = 1$ and $y = -1$ satisfy (4.3). However, we can resolve this "problem" by restricting the allowable values of y: If we specify in addition to (4.3) that $y \geq 0$, then (4.3) can be rewritten as

$$y = \sqrt{x}$$

and y is a function of x. All this should convince you that when a relationship between quantities is specified, determining whether or not this relationship can be described as one quantity being a function of another may depend on additional constraints being imposed, and sometimes this is needed in order to obtain a function. And, of course, the "real life" context may impose such or other constraints.

So let's look at the concept of a function more abstractly.[1] We'll also introduce some more handy notation going along with functions.

[1]In these lecture notes, we'll restrict ourselves to functions from numbers to numbers. More generally, one can consider functions from any set of objects to any other set of objects. E.g., one could define a function with domain the set of all Americans, which on the input "person" returns the value "eligible to vote" or "not eligible to vote". As long as we assume that for each American, exactly one of "eligible to vote" or "not eligible to vote" is true, this would also be a function, albeit admittedly much more abstract.

First of all, let's fix two sets of numbers X and Y; we'll call them the *domain* and the *codomain*[2] of our function f. We then use

$$f : X \to Y \tag{4.4}$$

to denote that f is a function with domain X and codomain Y. Next, for each *argument* x (i.e., each number x in the domain X), we have to specify *exactly one* number y (called the *value of f at x*) in the codomain Y; we'll then denote this number y as $f(x)$. To expand on our previous notation, we could write

$$\begin{aligned} f : X &\to Y \\ x &\mapsto f(x) \end{aligned} \tag{4.5}$$

So in this notation, the function $f(x) = 2x + 1$ could be written as

$$\begin{aligned} f : \mathbf{R} &\to \mathbf{R} \\ x &\mapsto 2x + 1 \end{aligned}$$

but we could also restrict the domain to the set of positive real numbers, say, and write

$$\begin{aligned} f : (0, \infty) &\to \mathbf{R} \\ x &\mapsto 2x + 1 \end{aligned}$$

or to the set of whole numbers and write

$$\begin{aligned} f : \mathbf{Z} &\to \mathbf{Z} \\ x &\mapsto 2x + 1 \end{aligned}$$

Note in this notation that the two arrows are slightly different: In the second line, we use the special arrow $\mapsto$ to indicate that x is "mapped to" y; so a function is sometimes also called a *map* or a *mapping*.

Often, in school mathematics, functions are given by simple formulas, such as $f(x) = 2x + 1$ or $f(x) = x^2$. But they do *not* have to be so simple! Here are three "perfectly reasonable" functions for which there

[2]While the notion of a domain is completely standard, the notion of a codomain is not, and often either no name is used for it at all, or it is sometimes called the "range", which, however, conflicts with our later notation.

are no "easy formulas":

$$f(x) = \begin{cases} 1 & \text{if } x \in \mathbf{Z} \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

$$g(x) = \begin{cases} x & \text{if } x \in \mathbf{Q} \\ 0 & \text{otherwise} \end{cases} \tag{4.7}$$

$$h(x) = \text{ the largest integer } n \leq x \tag{4.8}$$

Note that in each of (4.6)–(4.8), we clearly have a function in the sense in which we defined it at the beginning of the chapter, with domain the set of all real numbers: For each argument x, there is exactly one value y satisfying each of the three statements, respectively.

Another way to represent a function, apart from a "formula" or a description of a function as above, is in the form of a table. Table 4.1 shows some values for the functions f, g and h. The problem with a table is obviously that it can only show the values of a function at a very small number of values; and in the case of the function g, the table would not distinguish between g and the function $\hat{g}(x) = x$, for example, if we hadn't added the irrational number $\sqrt{2}$ to our table in spite of the fact that it looks a bit out of place!

x	-1	-0.5	0	0.5	1	$\sqrt{2}$
$f(x)$	1	0	1	0	1	0
$g(x)$	-1	-0.5	0	0.5	1	0
$h(x)$	-1	-0.5	0	0	1	1

TABLE 4.1. A table for the functions described in (4.6)–(4.8)

A third, and much more useful, way to represent a function is in the form of a graph. First of all, the *graph* of a function is the set of all points $(x, f(x))$ in the two-dimensional plane, where x varies over all arguments x in the domain of the function. Figures 4.1 and 4.2 show the graphs of f and h. The graph of g cannot be represented so easily: Since every irrational number x is arbitrarily close to a rational number x' (e.g., the decimal representation of x truncated at any fixed number of digits), we have $f(x) = 0$ and $f(x') = x'$ for arguments x and x' which can be arbitrarily close together, and that is not really possible to draw.

At the beginning of the section, we defined a function as mapping each argument x in the domain X to exactly one value y in the

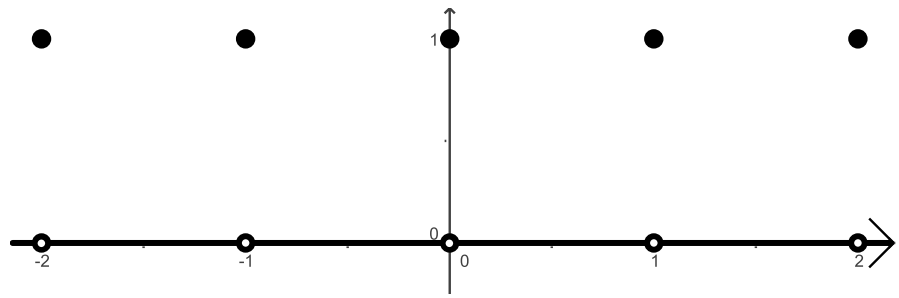

FIGURE 4.1. The graph of the function f from (4.6)

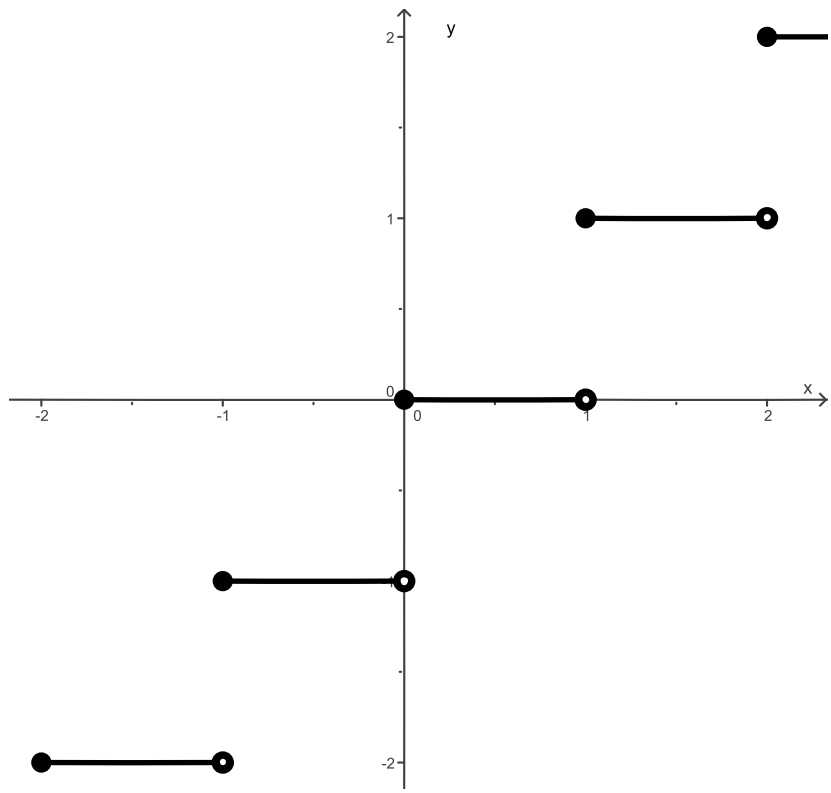

FIGURE 4.2. The graph of the function greatest integer function h from (4.8)

codomain Y. There is a visual way to check this property whether any given picture ("curve") in the plane is the graph of a function: To check whether for any argument x in the domain X, there is only one value y, we perform the "Vertical Line Test" by drawing a vertical

line at x and checking how often the curve intersects the vertical line. If the line and the curve intersect only once (for each argument x), then the curve "passes the Vertical Line Test" and is the graph of a function. This is because this unique point of intersection (x, y), say, tells us that the value of the function f at x is y. See Figure 4.3 for an example of a curve (described by the equation $x = y^2$) which does *not* pass the Vertical Line Test: The line given by $x = 1$ intersects the curve at the points $(1, 1)$ and $(1, -1)$. Algebraically, it's also easy to see that the curve has two parts, one described by $y = \sqrt{x}$ (the "upper half of the curve"), the other described by $y = -\sqrt{x}$ (the "lower half of the curve"); and both halves individually pass the Vertical Line Test, as they should, since they are the graphs of the two functions given by $y = \sqrt{x}$ and $y = -\sqrt{x}$, with domain the set of nonnegative real numbers.

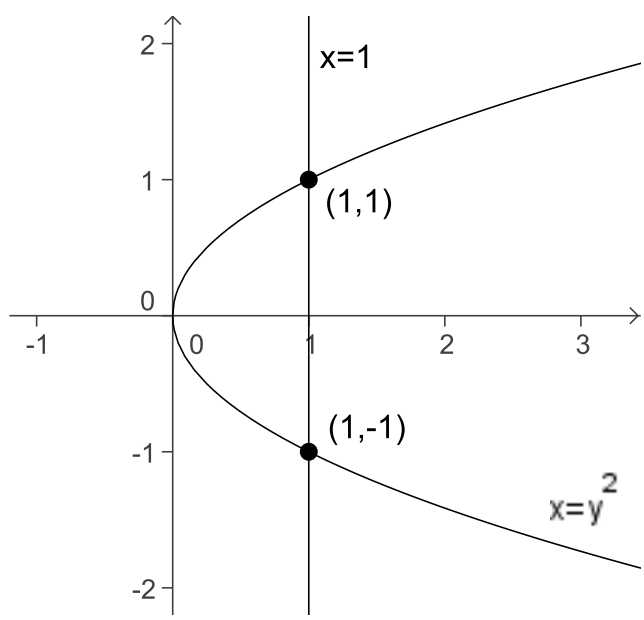

FIGURE 4.3. The Vertical Line Test for the curve $x = y^2$ with the line $x = 1$

4.2. Range of a Function, Onto and One-to-One Functions, and Inverse of a Function

We now discuss some properties of functions, leading up to the concept of "undoing" or "inverting" a function (i.e., given the value y, finding the argument x). Of course, this may not always be possible, as we will see.

4.2.1. Range of a Function, and Onto Functions. First of all, let the *range* of a function $f : X \to Y$ be the set of all y in the codomain Y of f which are of the form $f(x)$ for some argument x in the domain X of f; we will denote the range of f by $\mathrm{ran}(f)$. (Similarly, we will denote the domain of f by $\mathrm{dom}(f)$.)

Sometimes, such as in the example of the function $f : \mathbf{R} \to \mathbf{R}, x \mapsto 2x + 1$, the range of f is the same as the codomain: This is because any y in the codomain $\mathbf{R}$ can be written as $2x + 1$ by simply setting $x = \frac{y-1}{2}$. On the other hand, the range may consist of only a single number, such as in the case of the function $f : \mathbf{R} \to \mathbf{R}, x \mapsto 1$: Here only the number 1 is of the form $f(x)$, and in this case, we can choose x arbitrarily and will always obtain the value 1. In general, the range is a *subset* of the codomain, i.e., by definition of the codomain of a function $f : X \to Y$, any number of the form $f(x)$ for some x in the domain X must be in the codomain Y. There are many other possibilities for the range of a function: E.g., in the examples (4.6)–(4.8), we have that the range of f is $\{0, 1\}$, the range of g is $\mathbf{Q}$, and the range of h is $\mathbf{Z}$. This is easy to see from Figures 4.1 and 4.2 for the functions f and h. For the function g, we distinguish two cases for $g(x)$: If x is a rational number, then so is $g(x)$ (which just equals x), and any rational number y satisfies $y = g(y)$, so all rational numbers must be in the domain of g. On the other hand, if x is irrational, then $g(x)$ equals 0, which is rational; so the range of g is exactly the set of all rational numbers. In general, determining the range of a function is a very complicated problem!

We call a function $f : X \to Y$ *"onto"* (or *onto* Y, to be more precise) if the codomain and the range of the function f are the same. E.g., as remarked above, the function $f : \mathbf{R} \to \mathbf{R}, x \mapsto 2x + 1$ is onto, while the function $f : \mathbf{R} \to \mathbf{R}, x \mapsto 1$ is not. None of the functions in the examples (4.6)–(4.8) is onto as we saw above.

4.2.2. 1–1 Functions. Next, we call a function $f : X \to Y$ *one-to-one* (or *1–1* for short) if, for each value y in the codomain Y, there is at most one x in the domain X with $y = f(x)$. We can rephrase this in several equivalent ways:

(1) The function f is 1–1 if, for each value y in the range of f, there is exactly one x in the domain X with $y = f(x)$.
(2) The function f is 1–1 if, for all y in the codomain Y, whenever we have $y = f(x) = f(x')$ for arguments x and x' in the domain X, then actually $x = x'$.
(3) A visual way of checking whether a function is 1–1 is the Horizontal Line Test: Draw a horizontal line at the value y and

check how often the graph of the function intersects the vertical line. If the line and the graph intersect at most once (for each value y), then the graph "passes the Horizontal Line Test" and is the graph of a 1–1 function. This is because that point of intersection (x, y), say, tells us that the value of the function f equals y at at most one value x. (See Figure 4.4 for the example of the function $f : \mathbf{R} \to \mathbf{R}, x \mapsto x^2$ which fails the Horizontal Line Test, e.g., at the line $y = 1$, and thus is not 1–1. If we were to restrict the function f to the domain $[0, \infty)$ (i.e., the "right half" of the graph), or to the domain $(-\infty, 0]$ (i.e., the "left half" of the graph), then the function would pass the Horizontal Line Test and thus would be 1–1.)

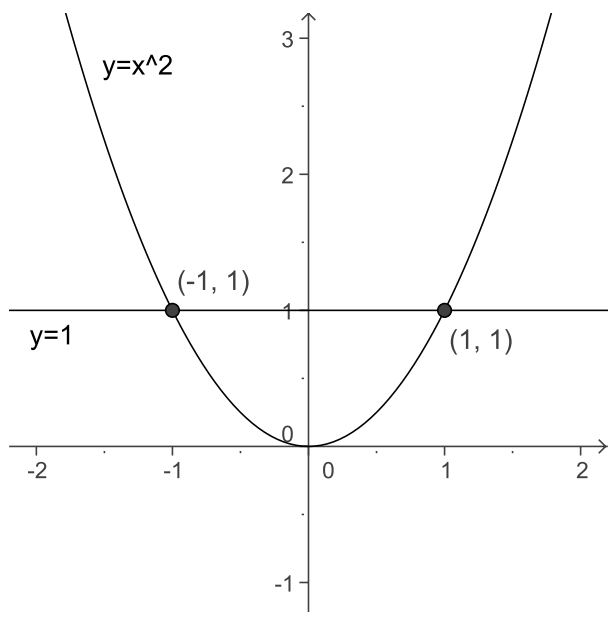

FIGURE 4.4. The Horizontal Line Test for the curve $y = x^2$ with the line $y = 1$

4.2.3. Inverse of a Function. Next we will investigate when we can "undo" or "invert" a function $f : X \to Y$ in the sense that we can find a function $g : Y \to X$ such that the statements $y = f(x)$ and $x = g(y)$ are equivalent. If such a function g exists, we will call it the *inverse function of* f and denote it by f^{-1}.[3]

[3]Note that the "exponent" -1 here does *not* denote that we take the "(-1)th power", i.e., $f^{-1}(x)$ does *not* mean $\frac{1}{f(x)}$! Later on, in trigonometry, when we

What are the conditions now under which the inverse function of a function $f : X \to Y$ exists? Note that we must ensure two things:

(1) For each y in the codomain Y of f, there is an argument x in the domain X of f such that $y = f(x)$; and
(2) for each y in the codomain Y of f, there is not more than one argument x in the domain X of f such that $y = f(x)$.

Now observe that condition (1) is exactly the condition that f is onto; and condition (2) is the condition that f is 1–1! These two conditions can be combined into one condition:

(3) For each y in the codomain Y of f, there is *exactly one* argument x in the domain X of f such that $y = f(x)$.

It should be clear that condition (3) is exactly the condition on the existence of an inverse function which we are looking for: Condition (1) is necessary since otherwise, for some y, we will not be able to define $f^{-1}(y)$. Condition (2) is necessary as well since otherwise, for some y, we will not know which value x to choose for $f^{-1}(y)$. On the other hand, condition (3) is also sufficient to ensure the existence of an inverse function $f^{-1} : Y \to X$: Given y in Y, condition (3) ensures the existence of exactly one x in X such that $f(x) = y$, and so it is both natural and inevitable to define $f^{-1}(y) = x$.

There is also a very nice visual way to find (the graph of) the inverse function of a function f: Recall that the graph of a function $f : X \to Y$ is the set of all pairs (x, y) such that $y = f(x)$, whereas the graph of the inverse function $f^{-1} : Y \to X$ is the set of all pairs (y, x) such that $x = f^{-1}(y)$, i.e., such that $y = f(x)$. So the graph of f^{-1} is obtained from the graph of f by simply interchanging the coordinates of each point (x, y)! Graphically, this corresponds to reflecting the graph of f along the "main diagonal" described by $y = x$; but, of course, this works only after we rename the variables from $x = f^{-1}(y)$ to $y = f^{-1}(x)$. Figures 4.5 and 4.6 show two examples.

We conclude with another example of our concepts for functions from this section: Let $X = \{0, 1, 2\}$ and $Y = \{1, 2, 3\}$ be two *finite* sets of numbers. Let f and g be two functions from X to Y, defined

---

commonly write expressions like $\sin^2 x$ in place of $(\sin x)^2$, this notation can lead to serious ambiguities, since we won't know how to interpret $\sin^{-1} x$ unless the context clarifies this, and I always recommend to use the special notation $\arcsin x$ for the inverse of the sine function in place of $\sin^{-1} x$; but we'll leave these intricacies aside for now.

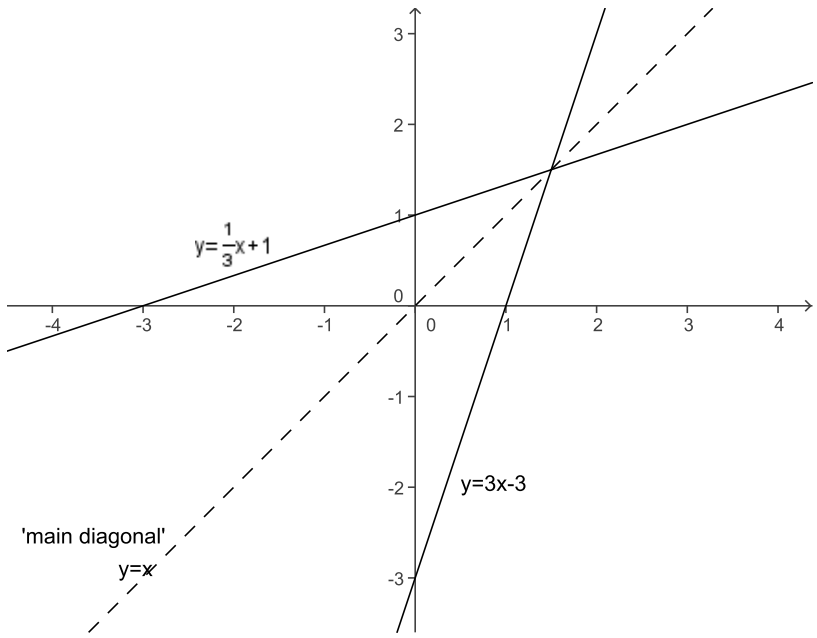

FIGURE 4.5. The graphs of the function $y = 3x - 3$ and its inverse function $y = \frac{1}{3}x + 1$

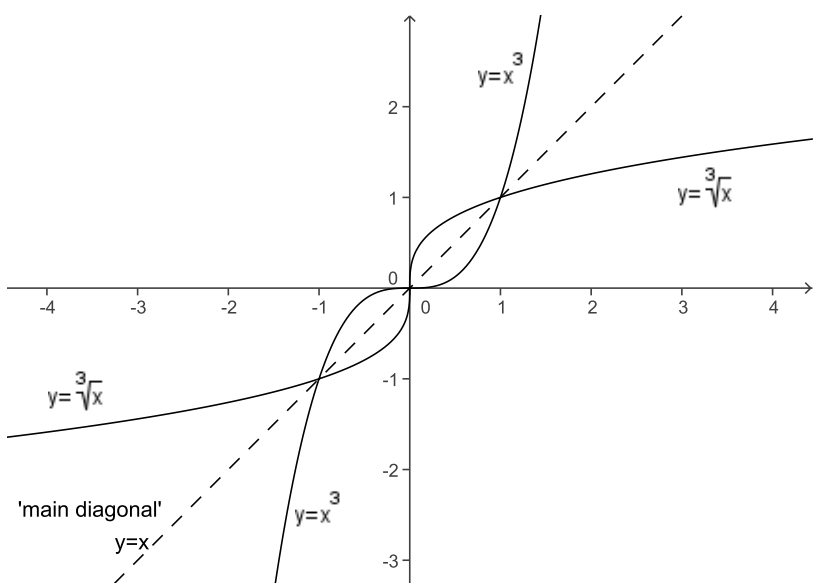

FIGURE 4.6. The graphs of the function $y = x^3$ and its inverse function $y = \sqrt[3]{x}$

by

$$f(0) = 1 \qquad g(0) = 1$$
$$f(1) = 3 \qquad g(1) = 3$$
$$f(2) = 3 \qquad g(2) = 2$$

We can represent f and g visually as shown in Figure 4.7.

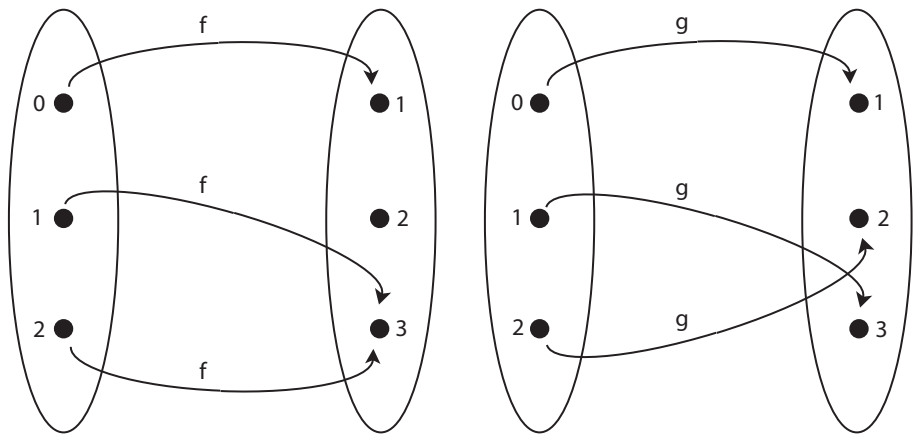

FIGURE 4.7. The finite functions f and g from $X = \{0, 1, 2\}$ to $Y = \{1, 2, 3\}$

It is now easy to see that f is neither onto nor 1–1: f is not onto since 2 is in the codomain but not in the range of f; and it is not 1–1 since $f(1) = f(2)$. It is therefore impossible to find an inverse function for f: Given the value 2 in the codomain Y of f, we can't find any x in the domain X of f with $f(x) = 2$; furthermore, for the value 3 in the codomain Y of f, there are two different x in the domain X of f with $f(x) = 3$. So it is not possible to find exactly one value for $f^{-1}(y)$ for either $y = 2$ or $y = 3$. On the other hand, for each y in Y, there is exactly one x in X such that $g(x) = y$, so we can define $g^{-1} : Y \to X$ with

$$g^{-1}(1) = 0$$
$$g^{-1}(2) = 2$$
$$g^{-1}(3) = 1$$

and g thus has an inverse function.

4.3. Some Functions Closely Related to Linear Functions

So far, we have mainly dealt with one kind of function, *linear functions*, which can be expressed as $f(x) = mx + b$ for fixed parameters m

and b. In this section, we will encounter some functions closely related to linear functions.

4.3.1. Step Functions. These are functions whose graph is horizontal except at certain points where it "jumps" from one value to another. We already saw an example of a step function in Figure 4.2; it is usually denoted by $y = \lfloor x \rfloor$, defined by

$$\lfloor x \rfloor = \text{ the largest integer } n \leq x \tag{4.8}$$

and called the *greatest integer function* (sometimes also the "floor function"). There is a similar function called the *least integer function* (or sometimes the "ceiling function"); it is defined by

$$\lceil x \rceil = \text{ the smallest integer } n \geq x \tag{4.9}$$

Its graph is shown in Figure 4.8.

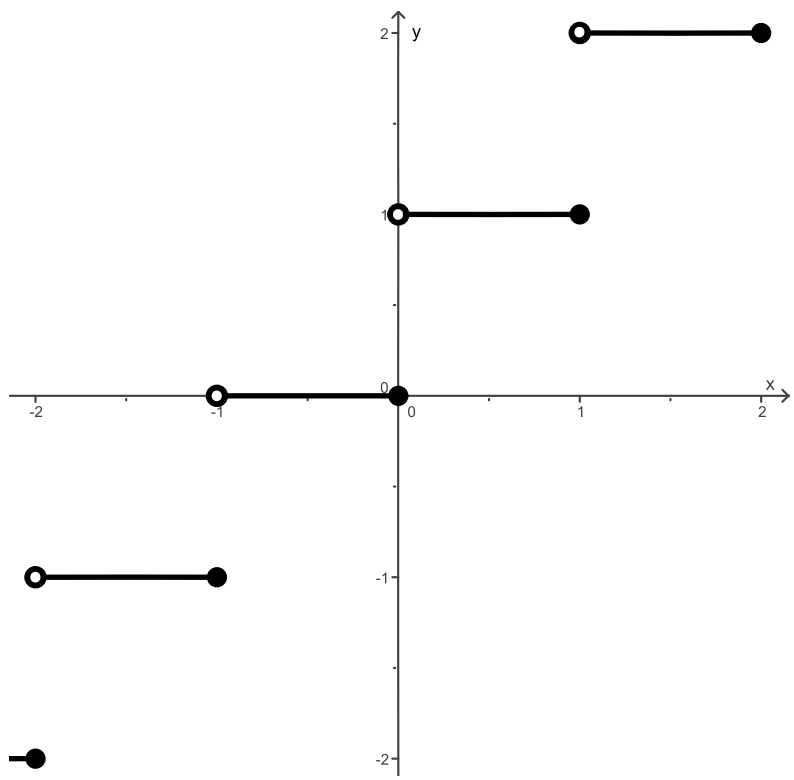

FIGURE 4.8. The graph of the least integer function from (4.9)

These two functions are particularly useful to computer scientists, who use them quite frequently. But they come up in "real life" situations quite early in school mathematics; in fact, we have already seen an example which really requires such a function if done carefully:

> "A pay phone charges 50¢ for the first minute and 10¢ for each additional minute of a long-distance call."

Before, we had restricted the domain of the function describing the cost of the phone call to the set of positive integers, so we defined this function as

$$\begin{aligned} f : \{1, 2, 3, \dots\} &\to \mathbf{R} \\ n &\mapsto 10n + 40 \end{aligned} \tag{4.10}$$

where n is the length of the phone call (in minutes) and $f(n)$ is its cost (in cents), and its graph was given in Figure 2.3. But this is really cheating, since phone calls can be of length any positive real (at least in principle), and then the phone company rounds up the length of the phone call to the next integer. So the cost function is really given by

$$\begin{aligned} g : (0, \infty) &\to \mathbf{R} \\ x &\mapsto 10 \cdot \lceil x \rceil + 40 \end{aligned} \tag{4.11}$$

The graph of the "real" cost function g is given in Figure 4.9.

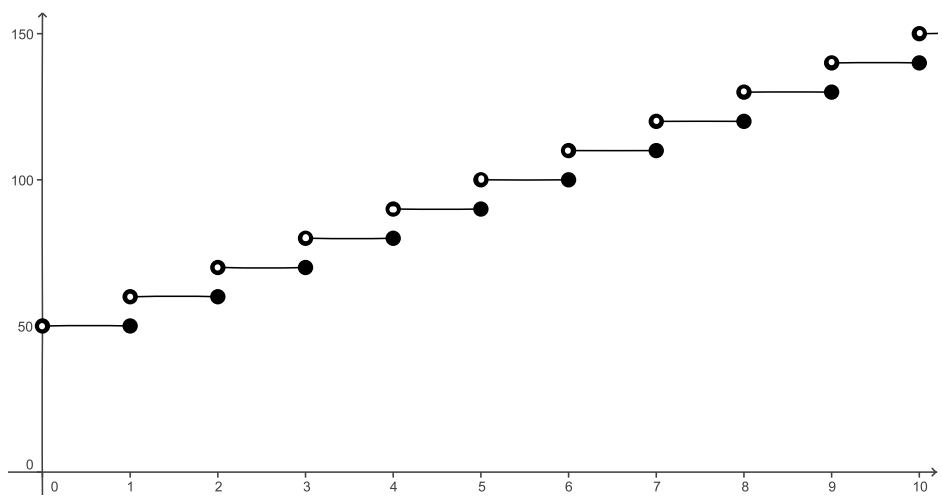

FIGURE 4.9. The cost of a phone call from (4.9)

What are the pros and cons of describing the cost function accurately as in (4.11) rather than just as $f(x) = 10x + 40$ (as in (2.6))? Certainly, the former is more accurate; it *is* the actual cost function as computed by the phone company. On the other hand, $f(x) = 10x + 40$ is a close approximation, which is much easier to work with: Unlike g, f is 1–1 and onto and thus has an inverse function; so we can easily compute the length of a 90-cent phone call, say, using f, or rather f^{-1}. If we tried this for an 85-cent phone call, say, we'd still get an "answer"

using f^{-1}, namely 4.5 minutes, and even though the answer would be wrong (there is no phone call costing exactly 85 cents!), it would tell us that the phone call would be around 4 or 5 minutes.

So the upshot is that a linear function is sometimes just a useful approximation to the real function, such as a step function, which is much harder to deal with. School mathematics is often happy to just deal with the approximation, but you as teachers should be aware of this distinction, and the reasons why one uses approximations, since it may confuse some of your students, and you should then be able to give at least a "roughly correct" answer.

4.3.2. Piecewise Linear Functions. Another very common kind of function one encounters in "real life" problems is a "piecewise linear" function: The graph of such a function is not a line, but consists of several line segments and is "bent" at the end points of these line segments, without any "jumps", however.

Here is an example:

> "John swims for half an hour at a constant speed of 2km/h, then jogs for an hour at a constant speed of 8km/h, and finally bikes for half an hour at a constant speed of 20km/h. Describe the distance traveled by John in terms of the time elapsed."

Let's let t (in hours) be the time elapsed since John starts, and s (in km) his distance traveled since then. Then the graph of s as a function of t is as shown in Figure 4.10.

Let's first describe this graph verbally: John first swims from time zero hours to time one-half hour at a speed of 2km/h, covering a distance of one kilometer. Then he jogs for one hour (from time one-half hour to time one-and-a-half hours) at a speed of 8km/h, covering an additional eight kilometers, for a total distance of nine kilometers for now. Finally, he bikes for half an hour (from time one-and-a-half hours to time two hours) at a speed of 20km/h, covering an additional ten kilometers, for final total distance of nineteen kilometers.

Let's translate this into an algebraic description of our function:

$$s(t) = \begin{cases} 2t & \text{if } 0 \le t \le \frac{1}{2} \\ 1 + 8(t - \frac{1}{2}) & \text{if } \frac{1}{2} \le t \le 1\frac{1}{2} \\ 9 + 20(t - 1\frac{1}{2}) & \text{if } 1\frac{1}{2} \le t \le 2 \end{cases} \tag{4.12}$$

So in the intervals $[0, \frac{1}{2}]$, $[\frac{1}{2}, 1\frac{1}{2}]$, and $[1\frac{1}{2}, 2]$, s behaves like the linear functions $s = 2t$, $s = 1 + 8(t - \frac{1}{2})$, and $s = 9 + 20(t - 1\frac{1}{2})$, respectively, as can be seen, e.g., from the point-slope formula of a linear function,

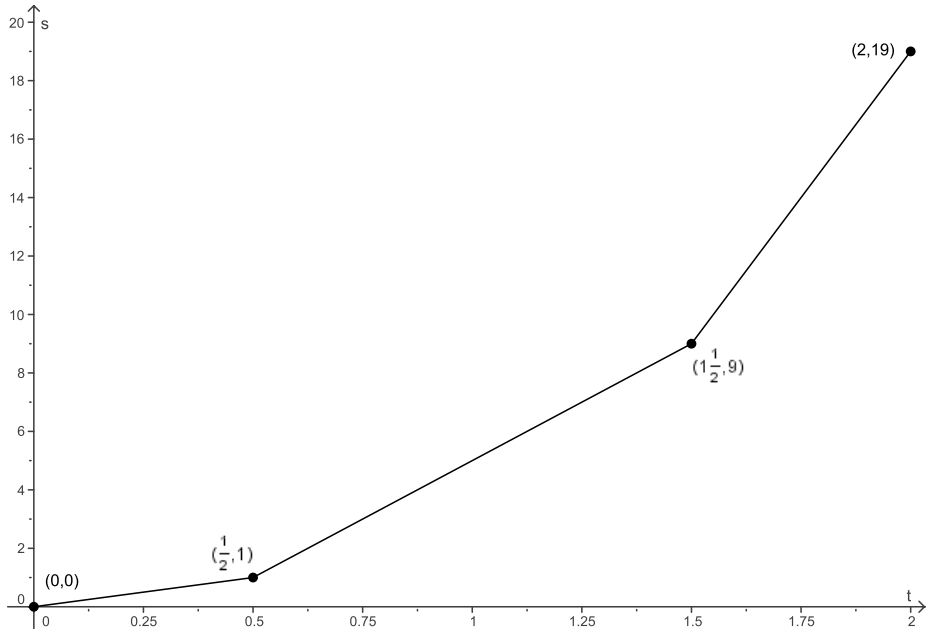

FIGURE 4.10. The distance s traveled by John terms of the time t elapsed

and as can be seen also from the graph shown in Figure 4.10. Note that at the points $t = \frac{1}{2}$, we can compute s both as $2 \cdot \frac{1}{2}$ and as $1 + 8(\frac{1}{2} - \frac{1}{2})$; so the function s "bends" but does not "jump" at $t = \frac{1}{2}$, and we are allowed to use $0 \leq t \leq \frac{1}{2}$ and $\frac{1}{2} \leq t \leq 1\frac{1}{2}$ in the first and second case of (4.12) even though the cases "overlap" at $t = \frac{1}{2}$. Similarly, at $t = 1\frac{1}{2}$, we can compute s both as $1 + 8(1\frac{1}{2} - \frac{1}{2})$ and as $9 + 20(1\frac{1}{2} - 1\frac{1}{2})$. The speeds at the various times corresponds to the slopes of the various line segments in the graph.

4.3.3. The Absolute Value Function. The absolute value function is a very important special case of a piecewise linear function. Informally, it gives the "distance" of a number x from the number 0 on the number line. Very informally, the absolute value of x is "x without the minus sign". As you probably already know, the absolute value of x is denoted by $|x|$. The graph of the absolute value function is shown in Figure 4.11.

The formal definition of the *absolute value* of x is given by

$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases} \tag{4.13}$$

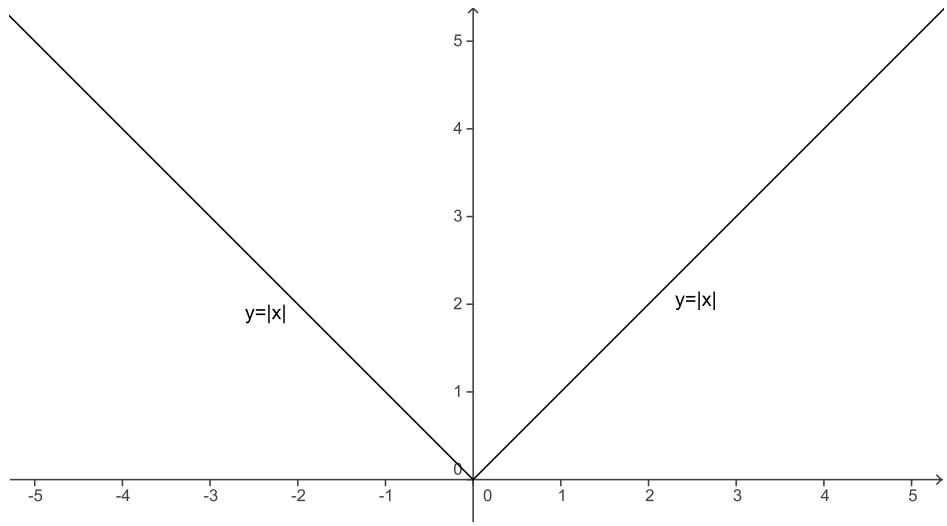

FIGURE 4.11. The graph of the absolute value function $y = |x|$

This simple and elegant definition is the source of endless confusion: If x is negative, then $-x$ is positive; so the second case of (4.13) always produces a *positive value*! Therefore, the absolute value of a number is always non-negative, one of the most important facts about the absolute value:

$$|x| \geq 0 \text{ for all } x \tag{4.14}$$

Before we discuss an example of how the absolute value function is used, let's state and discuss three important rules about this function, which hold for all real numbers a and all non-negative real numbers r:

$$|a| = r \text{ exactly when } a = r \text{ or } a = -r \tag{4.15}$$

$$|a| < r \text{ exactly when } -r < a < r \tag{4.16}$$

$$|a| > r \text{ exactly when } a < -r \text{ or } a > r \tag{4.17}$$

Why do properties (4.15)–(4.17) hold true? This is easy to "see intuitively" from the intuitive definition of $|x|$ as the "distance" of x from 0: The two numbers exactly at distance r from 0 are r and $-r$. (If $r = 0$, there is only one number here, of course.) The expression $|a| < r$ states that a is at a distance less than r from 0, which clearly holds exactly for the numbers in the interval $(-r, r)$. (This interval is empty if $r = 0$, as there are no numbers at a distance less than 0 from 0.) Finally, the expression $|a| > r$ covers all the cases not included in the first two cases, i.e., all numbers in the intervals $(-\infty, -r)$ and (r, ∞).

A second, more visual way to see why properties (4.15)–(4.17) hold true is to look at the graph in Figure 4.11: E.g., the expression $|a| < r$ means that we need to look at all values $< r$ on the y-axis, and the arguments x for which $|x| < r$ correspond to the points on the line segments from $(-r, r)$ to the origin, and from the origin to (r, r), not including the points $(-r, r)$ and (r, r) themselves.

A third, more algebraic way to see why properties (4.15)–(4.17) hold true is to analyze the statements on the left-hand sides in light of the formal definition of $|x|$ in (4.13): E.g., $|a| = r$ means that either $a \geq 0$ and so $|a| = a$, forcing $a = r$; or that $a < 0$ and so $|a| = -a$, forcing $-a = r$; this gives exactly the two cases on the right-hand side of (4.15). Similar arguments can be made for (4.16)–(4.17).

The main use of the absolute value lies in the fact that it does not only allow us to measure the distance of a number from 0 on the number line, but indeed the *distance on the number line* between any two numbers. Namely, the distance between numbers a and b is given by

$$|a - b|$$

The properties (4.15)–(4.17) now immediately translate to properties of the distance between numbers a and b (for any non-negative real number r)

$$|a - b| = r \text{ exactly when } a = b + r \text{ or } a = b - r \tag{4.18}$$

$$|a - b| < r \text{ exactly when } b - r < a < b + r \tag{4.19}$$

$$|a - b| > r \text{ exactly when } a < b - r \text{ or } a > b + r \tag{4.20}$$

since the right-hand sides of (4.18)–(4.20) are equivalent to the right-hand sides of (4.15)–(4.17) when we replace a by $a - b$.

A very common application of the absolute value is in estimating the accuracy of a measurement: We already saw that in "real life" applications, a length ℓ of 10 meters, say, really means a length of 10m$\pm$.5m, i.e., that

$$9.5 \leq \ell \leq 10.5 \tag{4.21}$$

(We actually wrote $9.5 \leq \ell < 10.5$ before, but let's ignore that fine point for the moment; in applications, it makes little difference.) Another way to express (4.21) would be

$$|\ell - 10| \leq .5$$

as can be seen from (4.18)–(4.19). Note that the negation of this statement, i.e., that the true length ℓ differs from 10m by more than half a

meter, is expressed by

$$|\ell - 10| > .5$$

or equivalently by

$$\ell < 9.5 \text{ or } \ell > 10.5$$

4.3.4. Direct and Inverse Variation. We already saw in Chapter 2 that often, the ratio (i.e., the quotient) between two quantities x and y, say, is a constant, which we'll call k:

$$k = \frac{y}{x} \tag{4.22}$$

This can be rewritten as

$$y = kx \tag{4.23}$$

In this case, we say that y *is proportional to* x or that y *varies directly with* x. An example of this from section 2.2 was constant speed v as the ratio of distance s traveled over time t:

$$v = \frac{s}{t} \tag{4.24}$$

At other times, we have that the *product* between two quantities x and y, say, is a constant, which we'll call k:

$$k = yx \tag{4.25}$$

This can again be rewritten as

$$y = \frac{k}{x} \tag{4.26}$$

In this case, we say that y *is inversely proportional to* x or that y *varies inversely with* x. The speed equation (4.24) also gives us an example of this if, instead of constant speed, we assume that the distance is constant and we want to study how the speed varies as we vary the time. (E.g., you may consider the travel over a fixed distance and thus consider the speed as depending on the time it takes you to travel this distance.) Figures 4.12 and 4.13 show the graphs of the functions in (4.23) and (4.26) for $k = 1$.

Sometimes, we want to study the relationship between more than two quantities, and then the concepts of direct and inverse variation are particularly useful. The so-called "Combined Gas Law" from physics, describing properties of gas in a sealed container, provides such an example:

$$\frac{PV}{T} = C \tag{4.27}$$

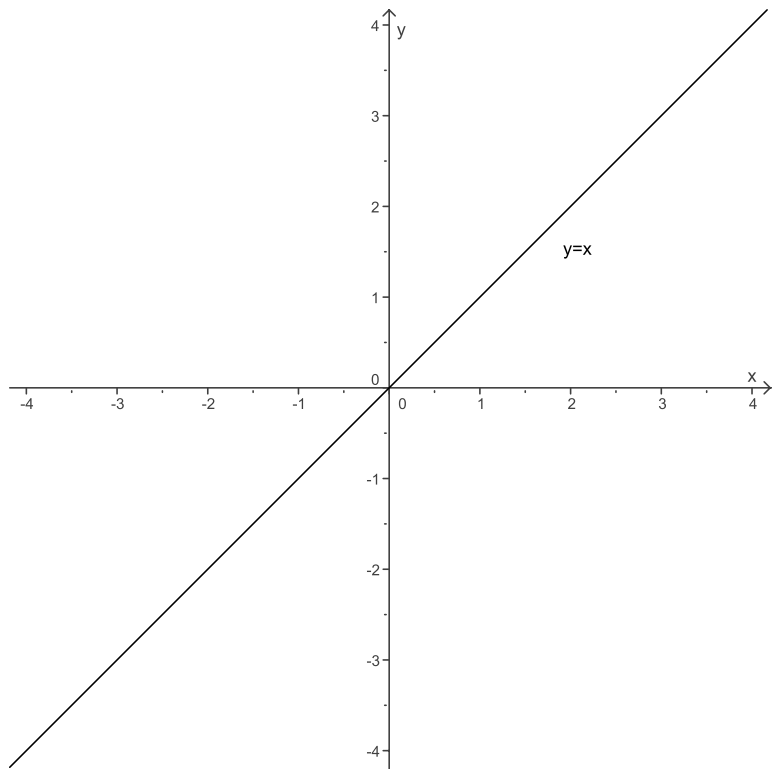

FIGURE 4.12. The graph of the function $y = x$

Here, P is the pressure (in pascal, say), V the volume (in m^3, say), T the absolute temperature (in degrees Kelvin), and C is a constant (which is fixed as long as we only consider a fixed amount of a particular kind of gas).

The Combined Gas Law can now be rewritten, and stated as saying the following:

(1) $P = \dfrac{CT}{V}$: The pressure varies directly with the temperature (at constant volume), and inversely with the volume (at constant temperature).

(2) $V = \dfrac{CT}{P}$: The volume varies directly with the temperature (at constant pressure), and inversely with the pressure (at constant temperature).

(3) $T = \dfrac{PV}{C}$: The temperature varies directly with the pressure (at constant volume), and also directly with the volume (at constant pressure).

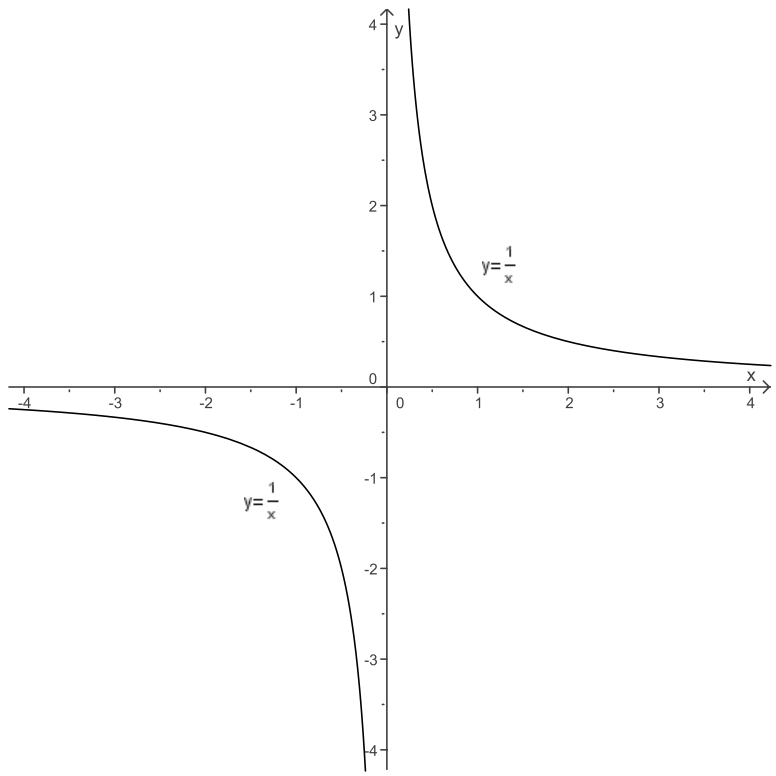

FIGURE 4.13. The graph of the function $y = \frac{1}{x}$

CHAPTER 5

Quadratic Functions, Equations and Inequalities

5.1. Introduction to Quadratic Functions

So far, the only functions we have considered are functions which are linear functions, or at least fairly close to linear functions. We will now turn to quadratic functions.

Let's start with an example. Consider the sequence of shapes in Figure 5.1.

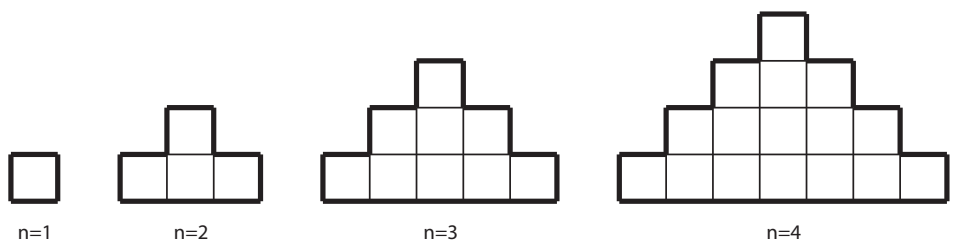

FIGURE 5.1. A sequence of shapes

There are several ways in which we could think of the $(n+1)$st shape being generated from the nth shape; here are two definitions leading to the same sequence:

(1) The $(n+1)$st shape is generated from the nth shape by adding a new row of squares below the last row of the nth shape, extending that last row one extra square to the left and one extra square to the right.
(2) The $(n+1)$st shape is generated from the nth shape by adding two new squares to the right of each row of the nth shape and then adding one new square above the middle of the new top row.

Whichever way you define the sequence of shapes, it is important to define it precisely: Simply letting the student "guess the pattern" is highly misleading, since there are many *different* ways to continue the "pattern". E.g., we could have a "pattern" which "stabilizes" at the 4th shape, i.e., the nth shape for all $n > 4$ is the same as the 4th shape.

Or, we could have a "pattern" which "oscillates" between the 1st and the 4th shape, i.e., the 5th shape is the same as the 3rd, the 6th shape is the same as the 2nd, the 7th shape is the same as the 1st, the 8th shape is the same as the 2nd again, etc. There is no reason why one pattern should be considered more "reasonable" than the others! This is why a careful definition of the intended pattern is essential.

Now let's look at the perimeter and the area of the nth shape in Figure 5.1 as defined above (assuming that the side length of each square is one unit length, and each square has area one unit square). Table 5.1 below shows these for the first five shapes.

number of shape	1	2	3	4	5
perimeter of shape	4	10	16	22	28
area of shape	1	4	9	16	25

TABLE 5.1. Perimeter and area of the shapes from Figure 5.1

We will now see that perimeter and area show a very different pattern:

Table 5.1 suggests that the perimeter increases by a fixed amount (namely, 6) each time. We can also see this, more precisely and for general n, from our definitions of the shapes: E.g., from the above definition (1) for the shapes, we see that each time we add to the pattern, we move the bottom edge down one unit; and then we add a square on each side of the new bottom row, resulting in an increase in perimeter by six edges of these two squares. We can also see this from the above definition (2) for the shapes: Each time we add to the pattern, we move all the edges on the right over by two units; and then we add four edges at the top (including for the new top square) and two edges at the bottom, resulting in an increase in perimeter by six edges.

Therefore, the perimeter (let's call it $P(n)$) of the nth shape is a linear function of n, with slope 6 and the graph containing the point $(1, 4)$ (since the 1st shape has perimeter 4). By the point-slope form for a line equation(see (2.11)), we arrive at the formula

$$P(n) = 6(n-1) + 4 = 6n - 2 \tag{5.1}$$

Here is another way of looking at finding the expression for this function: Let's define a new function (called the *first difference function*) by

$$P^1(n) = P(n+1) - P(n) \tag{5.2}$$

So this function P^1 tells us not the function P itself but by *how much* the function P increases from n to $n+1$. As we showed above, this increase is constant, namely, we have

$$P^1(n) = 6 \tag{5.3}$$

for all whole numbers $n > 0$. Note that (5.3) (together with (5.2)) does *determine* the function P; e.g., the function $P'(n) = 6n$ also satisfies both (5.3) and (5.2). But if we also specify that

$$P(1) = 4 \tag{5.4}$$

then we can compute the function P at all whole numbers $n > 0$ by starting with $n = 1$ in (5.4) and then repeatedly using that (5.3) and (5.2) give us

$$P(n+1) = P(n) + P^1(n) = P(n) + 6$$

to compute $P(2)$, $P(3)$, etc.[1]

We note that any function P (defined on the positive whole numbers) with constant first difference function P^1 must be linear: Suppose that $P^1(n) = c$ for all $n > 0$. Then we can compute

$$\begin{aligned}
P(2) &= P(1) + c \\
P(3) &= P(2) + c = P(1) + 2c \\
P(4) &= P(3) + c = \cdots = P(1) + 3c \\
&\cdots \\
P(n) &= P(n-1) + c = \cdots = P(1) + (n-1)c \\
&\cdots
\end{aligned}$$

On the other hand, the increase in area from one shape to the next is not constant: The increase from the 1st to the 2nd shape is 3, from the 2nd to the 3rd is 5, from the 3rd to the 4th is 7, etc. So, instead of the increase being *constant*, it appears from Table 5.1 that the increase *increases by a constant*, namely, 2. We can also see this more precisely from our definitions of the shapes: E.g., from the above definition (1) for the shapes, we see that each time we add to the pattern, we add two more squares to the shape than we did the last time (since the new bottom row is two units wider than the previous bottom row). And we can see this from Figure 5.2: The squares marked "new" (in regular font) are the same number as was added at the previous step; the two squares marked "*new*" (in slanted font) are added in addition to the

[1]Such a definition of a function is sometimes called a *definition by recursion*.

other new squares; so we do indeed add two more squares at each step than we did at the previous step.

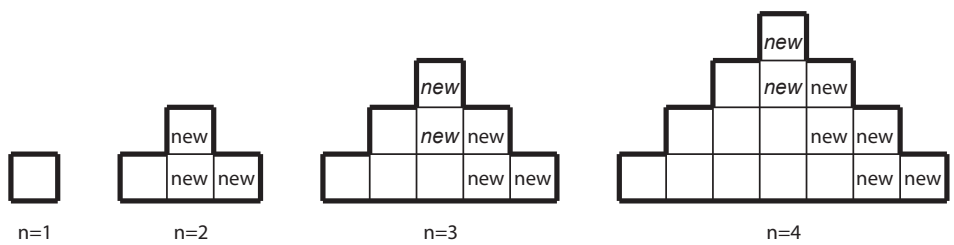

FIGURE 5.2. Adding new squares in our sequence of shapes

This results in the following equation for the first difference function A^1 of the area A:

$$A^1(n) = A(n+1) - A(n) = 2n + 1 \tag{5.5}$$

So, by the argument on the previous page, the first difference function $A^1(n) = A(n+1) - A(n)$ is a linear function, since it increases by 2 each time; namely, the so-called *second difference function*

$$A^2(n) = A^1(n+1) - A^1(n)$$

is constant, namely, 2.

Whenever the second difference function is constant, the first difference function must be linear, as we saw before; and whenever the first difference function is linear, then the function is quadratic, as we will now show.[2]

First of all, a *quadratic function* is a function of the form

$$A(n) = an^2 + bn + c \tag{5.6}$$

[2]If you happen to know some calculus, note the analogy: We consider here so-called discrete functions, i.e., functions defined on the positive integers, say, in an area called *discrete mathematics*. For such functions, the first difference function is the analogue of the first derivative of a continuous function, and the second difference function is the analogue of a continuous function. What we just said was that a discrete function is linear exactly when its first difference function is constant; and that a discrete function is quadratic exactly when its second difference function is constant. Of course, the derivatives do not exist for all continuous functions; but the difference functions of discrete functions always exist.

for parameters a, b and c. Now note that (5.6) satisfies

$$\begin{aligned} A^1(n) &= A(n+1) - A(n) \\ &= (a(n+1)^2 + b(n+1) + c) - (an^2 + bn + c) \\ &= (a(n+1)^2 - an^2) + (b(n+1) - bn) + (c - c) \\ &= (an^2 + 2an + a - an^2) + (bn + b - bn) \\ &= (2a)n + (a+b) \end{aligned} \tag{5.7}$$

and so $A^1(n)$ is a linear function of n. But we have

$$A(n+1) = A(n) + (A(n+1) - A(n)) = A(n) + A^1(n)$$

and so having $A(1)$ and a formula for $A^1(n)$ completely determines the function A.

Comparing (5.5) and (5.7), we now have

$$2n + 1 = (2a)n + (a+b)$$

and so

$$\begin{aligned} 2 &= 2a \\ 1 &= a + b \end{aligned}$$

which solves to

$$\begin{aligned} a &= 1 \\ b &= 0 \end{aligned}$$

Now we still have to find c for (5.6). Setting $n = 1$, we also have

$$A(1) = 1 = 1^2 + c$$

and therefore $c = 0$. Thus the area of the nth shape in Figure 5.1 is

$$A(n) = n^2 \tag{5.8}$$

Of course, you probably "guessed" the formula for the area all along from Table 5.1. However, I hope you found the "formal" way of deriving the formula above instructive, since it works for *any* sequence of values with constant second difference; we'll work out a second example below to show how this works in a more complicated case. But let's first prove the formula (5.8) for the area of the nth shape geometrically: Figure 5.3 shows how to move some of the "old" squares to "new" places in such a way that the nth shape becomes a square of side length n and thus of area n^2.

Now, as promised, let's slightly vary the problem: Consider the shapes in Figure 5.4. It is now no longer so easy to guess a formula for

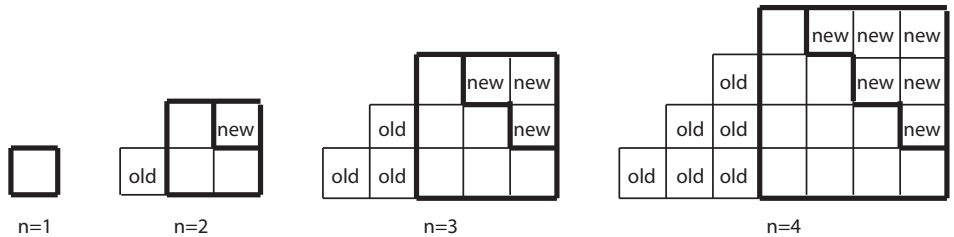

FIGURE 5.3. Computing the area of the nth shape in our sequence of shapes from Figure 5.1

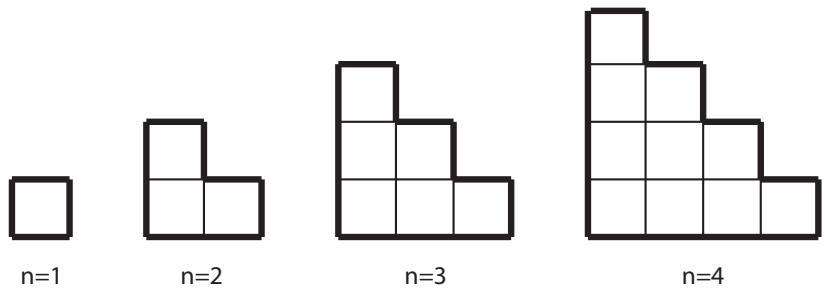

FIGURE 5.4. A new sequence of shapes

the area of the nth shape. But it's not hard to see that the areas of these shapes satisfy

$$A(n+1) - A(n) = n + 1 \tag{5.9}$$

and so the second difference is constant 1. Comparing (5.9) with (5.7) above, we arrive at

$$n + 1 = (2a)n + (a + b)$$

and so

$$\begin{aligned} 1 &= 2a \\ 1 &= a + b \end{aligned}$$

which solves to

$$\begin{aligned} a &= \frac{1}{2} \\ b &= \frac{1}{2} \end{aligned}$$

Now we still need to find c. Setting $n = 1$ gives

$$A(1) = 1 = \tfrac{1}{2} \cdot 1^2 + \tfrac{1}{2} + c$$

and therefore again $c = 0$. Thus the area of the nth shape in Figure 5.1 is

$$A(n) = \tfrac{1}{2}n^2 + \tfrac{1}{2}n \tag{5.10}$$

EXERCISE 5.1. Find a formula for the perimeter of the shapes in Figure 5.4, and prove both it and the area formula (5.10) geometrically in at least two different ways.

5.2. Solving Quadratic Equations

We call an equation a *quadratic equation* (in an unknown x) if it is of the form

$$ax^2 + bx + c = 0 \tag{5.11}$$

for some parameters a, b, c in $\mathbf{R}$. We will also assume from now on that $a \neq 0$ since otherwise (5.11) reduces to a linear equation. Note that a more general equation

$$ax^2 + bx + c = a'x^2 + b'x + c'$$

can easily be transformed into the format (5.11) by "subtracting the right-hand side to the left"; so we can restrict ourselves in general to equations of the form (5.11).

We will now consider increasingly sophisticated (and more complicated) techniques to solve a quadratic equation, starting with a review of the binomial laws, which will prove very useful.

5.2.1. Reviewing the Binomial Laws. The *binomial laws* (for algebraic expressions a, b and c) state that

$$(a + b)^2 = a^2 + 2ab + b^2 \tag{5.12}$$

$$(a - b)^2 = a^2 - 2ab + b^2 \tag{5.13}$$

$$(a + b)(a - b) = a^2 - b^2 \tag{5.14}$$

and are simple but very useful consequences of the laws of arithmetic from Chapter 1.

We will prove them in two different ways, algebraically from the laws of arithmetic, and geometrically using models for addition and multiplication.

Let's start with an algebraic proof of (5.12):

$$\begin{aligned}
(a + b)^2 &= (a + b)(a + b) && \text{(by definition of squaring)}\\
&= a(a + b) + b(a + b) && \text{(by distributive law)}\\
&= (a^2 + ab) + (ba + b^2) && \text{(by distributive law again)}\\
&= a^2 + 2ab + b^2 && \text{(by comm./assoc./distrib. laws)}
\end{aligned}$$

For a geometric proof of (5.12) (for $a, b > 0$), refer to Figure 5.5. The area of the big square, which has side length $a+b$, is $(a+b)^2$, and is the sum of the areas of the two smaller squares and the two rectangles, which have areas a^2, b^2, ba and ab, respectively. The binomial law (5.12) now follows immediately.

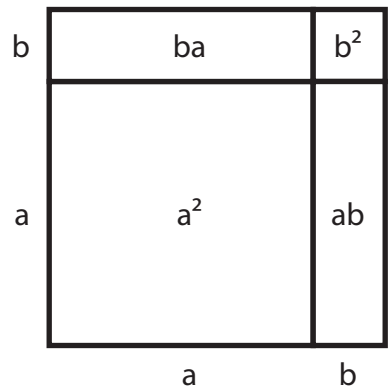

FIGURE 5.5. A geometric proof (for $a, b > 0$) showing that $(a+b)^2 = a^2 + 2ab + b^2$

EXERCISE 5.2. Give similar algebraic and geometric proofs of the binomials laws (5.13)–(5.14). (For the geometric proofs, assume that $0 < b < a$ to make geometric sense. Explain why this assumption is desirable for the geometric proofs.)

5.2.2. Solving Quadratic Equations by Factoring. Let's start with an example:

$$x^2 + 2x - 3 = 0 \tag{5.15}$$

Since $x^2 + 2x - 3 = (x-1)(x+3)$ (by using the distributive property repeatedly!), this can be rewritten as

$$(x-1)(x+3) = 0 \tag{5.16}$$

Of course, it's not always so easy to see this immediately, and in fact it is often impossible to "see" such a "factoring" of the left-hand side of a quadratic equation at all. There is one useful fact from advanced algebra[3] which often comes in handy: Suppose we have an equation of

[3]Here I am using the word "algebra" not in the sense of school algebra, but in the sense in which research mathematicians use it. The proof of this fact is typically covered in an advanced undergraduate course on "Modern Algebra".

the form $x^2 + bx + c = 0$ (i.e., the parameter a equals 1, or, put differently, x^2 is not preceded by a number) and the parameters b and c are integers, then all the solutions are (positive or negative) integers which divide the parameter c, or else necessarily involve square roots. So in the above example (5.15), we know right away, without even factoring, that the solutions to (5.15) either involve square roots or are among the numbers $1, -1, 3, -3$ (which are the only integers dividing 3).

Now we use a simple fact about the real numbers: Multiplying two nonzero numbers always produces a nonzero number:

$$A \neq 0 \text{ and } B \neq 0 \text{ implies } A \cdot B \neq 0 \tag{5.17}$$

Equivalently, if a product of two numbers is zero, then one of the factors has to be zero:

$$A \cdot B = 0 \text{ implies } A = 0 \text{ or } B = 0 \tag{5.18}$$

Both (5.17) and (5.18) hold for all numbers A and B. And since the converse of these is trivial (a product where one factor is zero must be zero), we actually have

$$A \cdot B = 0 \text{ holds exactly when } A = 0 \text{ or } B = 0 \tag{5.19}$$

Now let's apply (5.19) to (5.16):

$$(x+1)(x-3) = 0 \text{ holds exactly when } x + 1 = 0 \text{ or } x - 3 = 0$$

which shows that the equation (5.15) has exactly the two solutions -1 and 3.

More generally, and again using (5.19), the solution(s) to a quadratic equation of the form

$$(x + A)(x + B) = 0$$

are $-A$ and $-B$. (Of course, if $A = B$ then there is only one solution here.)

Let's talk about another very simple way to solve quadratic equations of another special form next.

5.2.3. Solving Quadratic Equations by Taking the Square Root. Let's again start with an example: Consider the quadratic equation

$$x^2 + 2x = 3$$

Since $x^2 + 2x + 1 = (x + 1)^2$, this can be rewritten as

$$(x + 1)^2 = 4$$

after adding 1 to both sides. (Of course, it's not always so easy to see this immediately, but we will learn a technique in subsection 5.2.4 which shows that this can always be done.) The point of this rewriting

is that now we can "take the square root" on both sides and we arrive at

$$x+1 = \pm 2 \quad \text{(which is an abbreviation for "}x+1=2 \text{ or } x+1=-2\text{")}$$

(Don't forget the $\pm$ here! More about that in a minute.) Subtracting 1 on both sides then yields

$$x = -1 \pm 2 \quad \text{(which is an abbreviation for "}x=1 \text{ or } x=-3\text{")}$$

Clearly, "taking the square root" here did the trick, but why was that justified? This step, and the $\pm$ part of it, is a frequent source of student errors, so let's look at it more closely: What we really did was to take the following steps, each resulting in an equivalent statement:

$$(x+1)^2 = 4 \tag{5.20}$$

$$|x+1|^2 = |2|^2 \tag{5.21}$$

$$|x+1| = |2| \tag{5.22}$$

$$x+1 = \pm 2 \qquad \text{(i.e., } x+1=2 \text{ or } x+1=-2\text{)} \tag{5.23}$$

Let's analyze each step carefully: Going from (5.20) to (5.21) uses that $4 = 2^2$, that $2 = |2|$, and the identity

$$A^2 = |A|^2$$

which holds for all algebraic expressions A and follows immediately from the definition (4.13) of the absolute value. Going from (5.21) to (5.22) is a new operation on equations for us: If both sides of an equation $A = B$ (for algebraic expressions A and B) are non-negative, then the equations

$$A = B \qquad \text{and} \qquad A^2 = B^2$$

have the same solutions. Finally, the step from (5.22) to (5.23) follows immediately from (4.15). Since this sequence of three steps is used quite frequently, we state it as a proposition:

Proposition 5.3. *The solutions to the equation*

$$A^2 = B^2$$

(for algebraic expressions A and B) are the same as the solutions to the statement

$$A = B \text{ or } A = -B$$

The proof of Proposition 5.3 is the same the proof for the special case above: The following four statements are equivalent, by the same

reasoning as for (5.20)–(5.23):

$$A^2 = B^2$$
$$|A|^2 = |B|^2$$
$$|A| = |B|$$
$$A = B \text{ or } A = -B$$

Proposition 5.3 now gives us a general procedure to solve quadratic equations of the form $(x + A)^2 = B$ (for any numbers A and B). But there is a small wrinkle we haven't considered yet: What if B is negative? Since the square $(x + A)^2$ is always non-negative, there cannot be any solution (at least in the real numbers[4]) in this case!

So we'll assume in addition that $B \geq 0$. Then the solution(s) to the quadratic equation $(x + A)^2 = B$ are

$$x = \sqrt{B} - A \text{ and } x = -\sqrt{B} - A$$

Of course, there will be two different solutions if B is positive, and only one solution (namely, $-A$) if $B = 0$. This shows a particularly simple way to solve quadratic equations, and we'll see in subsection 5.2.4 that we can always reduce the solution of a quadratic equation to this case.

5.2.4. Solving Quadratic Equations by Completing the Square. In subsection 5.2.2, we saw that sometimes a quadratic equation $x^2 + bx + c = 0$ can be solved very efficiently, namely, whenever we "see" how to factor the "polynomial" $x^2 + bx + c$ as a product $(x + A)(x + B)$ for some numbers A and B. But if the solutions to a quadratic equation are too complicated (e.g., if they involve square roots), then we need another way to solve the equation, which will always lead to a solution. Subsection 5.2.3 already gave a hint as to how we might proceed.

First of all, we can always reduce solving an arbitrary quadratic equation to an equation of the form[5]

$$x^2 + px + q = 0 \quad (\text{e.g., } x^2 + 2x - 3 = 0) \tag{5.24}$$

simply by dividing the original quadratic equation by the number "before" x^2 (which is usually called the *coefficient* of x^2).

Next, we can rewrite (5.24) as

$$x^2 + px = -q \quad (\text{e.g., } x^2 + 2x = -(-3) = 3) \tag{5.25}$$

[4]In subsection 5.3.2, we'll briefly cover the complex numbers, in which we can solve the equation $(x + A)^2 = B$ even when B is negative.

[5]We will always accompany our general considerations in this subsection by an example to make things clearer.

by subtracting q on both sides.

Now we can add $\frac{1}{4}p^2$ to both sides of (5.25) to arrive at

$$x^2 + px + \tfrac{1}{4}p^2 = \tfrac{1}{4}p^2 - q \quad (\text{e.g., } x^2 + 2x + 1 = 1 + 3 = 4) \tag{5.26}$$

and we are in the case of subsection 5.2.3, since we can now rewrite (5.26) as

$$\left(x + \tfrac{1}{2}p\right)^2 = \tfrac{1}{4}p^2 - q \quad (\text{e.g., } (x+1)^2 = 4) \tag{5.27}$$

(See Figure 5.6 for a geometric picture of how completing the square works for the left-hand side of the equation: The area corresponding to px is split in half; each half is put along a side of a square of area x^2. Now, in order to get a square again, we need to add a square of area $\frac{1}{4}p^2$ in the corner, to arrive at a square of area $(x + \frac{1}{2}p)^2$. Of course, this picture assumes $p > 0$; but a similar picture can be drawn for the case $p < 0$.)

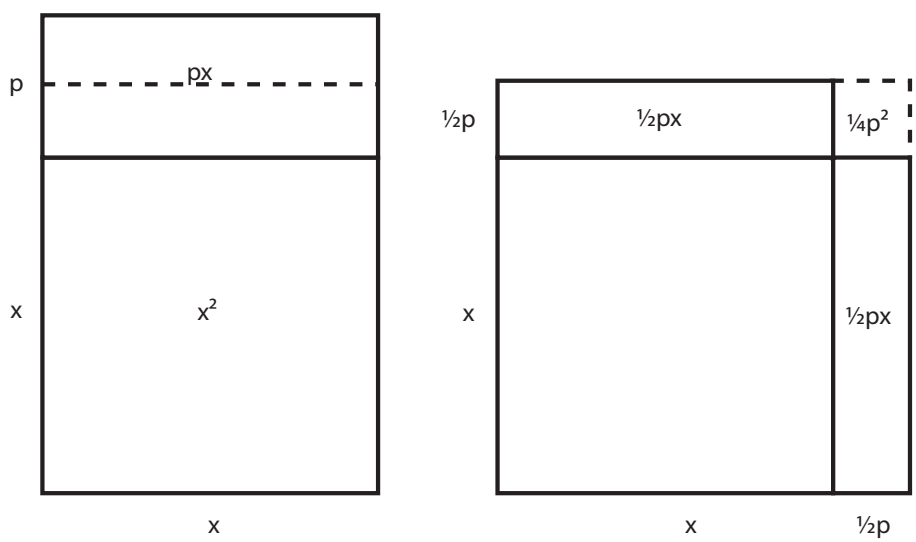

FIGURE 5.6. Completing the square geometrically

As in subsection 5.2.3, we can now solve (5.24) as

$$x + \tfrac{1}{2}p = \pm\sqrt{\tfrac{1}{4}p^2 - q} \quad (\text{e.g., } x + 1 = \pm 2) \tag{5.28}$$

and so

$$x = -\tfrac{1}{2}p \pm \sqrt{\tfrac{1}{4}p^2 - q} \quad (\text{e.g., } x = -1 \pm 2) \tag{5.29}$$

The main "trick" in this way of solving the quadratic equation is in going from (5.25) to (5.26); this step is often calling *completing the*

square since we are adding just the right number in order to make the left-hand side into an expression which can be written as a square, as shown geometrically in Figure 5.6.

5.2.5. Solving Quadratic Equations by the Quadratic Formula. The format of (5.29) above already suggests that one could solve a general quadratic equation (with parameters in place of numbers) by simply finding a general formula which gives the solution(s) to the equation. Such a formula does indeed exist, and we will now deduce it in its most general form. ((5.29) is the formula for the special form of the quadratic equation in (5.24), i.e., when the coefficient of x^2 is 1.)

We start with the general quadratic equation[6]

$$ax^2 + bx + c = 0 \quad (\text{e.g., } 2x^2 + 7x + 5 = 0) \tag{5.30}$$

(and still assume that $a \neq 0$).

First, we can rewrite (5.30) as

$$ax^2 + bx = -c \quad (\text{e.g., } 2x^2 + 7x = -5) \tag{5.31}$$

by subtracting c on both sides, and then as

$$x^2 + \frac{b}{a}x = -\frac{c}{a} \quad (\text{e.g., } x^2 + \tfrac{7}{2}x = -\tfrac{5}{2}) \tag{5.32}$$

by dividing by a on both sides.

Now we can add $\left(\frac{b}{2a}\right)^2$ to both sides of (5.32) to arrive at

$$x^2 + 2\frac{b}{2a}x + \left(\frac{b}{2a}\right)^2 = \left(\frac{b}{2a}\right)^2 - \frac{c}{a} = \frac{b^2 - 4ac}{4a^2} \tag{5.33}$$
$$(\text{e.g., } x^2 + \tfrac{7}{2}x + \left(\tfrac{7}{4}\right)^2 = \left(\tfrac{7}{4}\right)^2 - \tfrac{5}{2} = \tfrac{49-40}{16} = \tfrac{9}{16})$$

and we are in the case of subsection 5.2.3, since we can now rewrite (5.33) as

$$\left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2} \quad (\text{e.g., } \left(x + \tfrac{7}{4}\right)^2 = \tfrac{9}{16}) \tag{5.34}$$

As in subsection 5.2.3, we can now solve (5.30) as

$$x + \frac{b}{2a} = \pm\frac{\sqrt{b^2 - 4ac}}{2a} \quad (\text{e.g., } x + \tfrac{7}{4} = \pm\tfrac{3}{4}) \tag{5.35}$$

[6]Again, we will accompany our general considerations in this subsection by an example to make things clearer.

and so arrive at the *quadratic formula*

$$x = -\frac{b}{2a} \pm \frac{\sqrt{b^2 - 4ac}}{2a} \tag{5.36}$$

(e.g., $x = -\frac{7}{4} \pm \frac{3}{4}$, i.e., $x = -1$ or $x = -\frac{5}{2}$)

for the *solution to the general quadratic equation*

$$ax^2 + bx + c = 0 \tag{5.30}$$

We see from the quadratic formula (5.36) that the quantity $b^2 - 4ac$ is crucial; it is called the *discriminant* of the quadratic equation and determines the number of solutions (in the real numbers): If the discriminant is positive, then (5.36) gives two solutions. If the discriminant is 0, then (5.36) gives only one solution (namely, $-\frac{b}{2a}$). Finally, if the discriminant is negative, then (5.36) gives no solution, which makes sense if we look back at (5.34): If the discriminant is negative, then the right-hand side of (5.34) is negative (since the denominator of the right-hand side is a square and thus positive), whereas the left-hand side of (5.34) is a square and thus non-negative.

To summarize, there are different ways to solve a quadratic equation, and there is not one "best" way. While the technique of completing the square and the quadratic formula will always work, the techniques we discussed earlier are sometimes more efficient. Furthermore, the quadratic formula is rather complicated, and it's easy to memorize it wrong; so at the very least, one should remember along with the quadratic formula also how to derive it so that one can quickly reconstruct it.

5.2.6. The number of real solutions of a quadratic equation. We already saw in the last two subsections that a quadratic equation

$$ax^2 + bx + c = 0 \tag{5.30}$$

has at most two solutions in the real numbers; and we will see this again graphically in section 5.4. Algebraically, this fact depended on Proposition 5.3, which states that for all algebraic expressions A and B, the equation $A^2 = B^2$ holds exactly when $A = B$ or $A = -B$ holds.

There is another proof of the fact that any quadratic equation can have at most two real solutions, which uses factoring along the lines of subsection 5.2.2. Let's start with an example: Consider the equation

$$x^2 + 3x + 1 = 0 \tag{5.37}$$

From the quadratic formula, we can see that one solution of (5.37) is $-\frac{3}{2}+\frac{1}{2}\sqrt{5}$. (Of course, the "other" solution from the formula is $-\frac{3}{2}-\frac{1}{2}\sqrt{5}$, but let's ignore that for now.) We will now "divide" the expression x^2+3x+1 by the expression $x+\frac{3}{2}-\frac{1}{2}\sqrt{5}$ (i.e., the expression "$x-$ solution", for a reason we'll see later), similarly to the way long division works for whole numbers, except that now the "digits" of our expressions are not numbers between 0 and 9 (as in usual long division) but expressions of the form ax^2, bx or c where a, b and c are numbers. This procedure is called *polynomial division* and works in our example as shown in Table 5.2.

$$
\begin{array}{r|rcrcr}
 & & & x & + & (\frac{3}{2}+\frac{1}{2}\sqrt{5}) \\
\hline
x+(\frac{3}{2}-\frac{1}{2}\sqrt{5}) & x^2 & + & 3x & + & 1 \\
 & -\ (x^2 & + & (\frac{3}{2}+\frac{1}{2}\sqrt{5})x) & & \\
\cline{2-4}
 & & & (\frac{3}{2}-\frac{1}{2}\sqrt{5})x & + & 1 \\
 & & - & ((\frac{3}{2}-\frac{1}{2}\sqrt{5})x & + & 1) \\
\cline{3-6}
 & & & & & 0
\end{array}
$$

TABLE 5.2. Dividing $(x^2+3x+1)\div(x+\frac{3}{2}-\frac{1}{2}\sqrt{5})$ by polynomial division

Let's review this process of polynomial division carefully: We want to divide x^2+3x+1 by $x+(\frac{3}{2}-\frac{1}{2}\sqrt{5})$. Just as in long division, we do it "digit" by "digit", except that instead of "digits", we now use the coefficients: In x^2+3x+1 , the coefficient 1 of x^2 corresponds to the "hundreds digit", the coefficient 3 of x corresponds to the "tens digit", and the constant coefficient 1 corresponds to the "ones digit"; in $x+(\frac{3}{2}+\frac{1}{2}\sqrt{5})$, the coefficient 1 of x corresponds to the "tens digit", and the constant coefficient $\frac{3}{2}+\frac{1}{2}\sqrt{5}$ corresponds to the "ones digit".

We first check "how often x goes into" x^2 and find that it "goes x times". So we multiply $x+(\frac{3}{2}+\frac{1}{2}\sqrt{5})$ by x and subtract the result from x^2+3x+1, leaving us with $(\frac{3}{2}-\frac{1}{2}\sqrt{5})x$. We now "take down" the coefficient 1 and check again "how often x goes into" $(\frac{3}{2}-\frac{1}{2}\sqrt{5})x$; we find it "goes $(\frac{3}{2}-\frac{1}{2}\sqrt{5})$ times", multiply $x+(\frac{3}{2}+\frac{1}{2}\sqrt{5})$ by $\frac{3}{2}-\frac{1}{2}\sqrt{5}$, and subtract the result from $(\frac{3}{2}-\frac{1}{2}\sqrt{5})x+1$, leaving us with remainder 0. So x^2+3x+1 divides evenly into $x+(\frac{3}{2}+\frac{1}{2}\sqrt{5})$ (i.e., there is no remainder), and the quotient is $x+(\frac{3}{2}-\frac{1}{2}\sqrt{5})$. We can write this also

as

$$\frac{x^2+3x+1}{x+\left(\frac{3}{2}+\frac{1}{2}\sqrt{5}\right)} = x + \left(\frac{3}{2}-\frac{1}{2}\sqrt{5}\right) \qquad \text{or}$$

$$x^2+3x+1 = \left(x+\left(\frac{3}{2}-\frac{1}{2}\sqrt{5}\right)\right)\left(x+\left(\frac{3}{2}+\frac{1}{2}\sqrt{5}\right)\right)$$

So why all this work? We can now use (5.18): The solutions x to the equation

$$x^2+3x+1=0 \tag{5.37}$$

are exactly the solutions to the statement

$$x+\left(\frac{3}{2}+\frac{1}{2}\sqrt{5}\right)=0 \text{ or } x+\left(\frac{3}{2}-\frac{1}{2}\sqrt{5}\right)=0$$

Thus equation (5.37) has exactly two solutions, namely, $-\frac{3}{2}-\frac{1}{2}\sqrt{5}$ and $-\frac{3}{2}+\frac{1}{2}\sqrt{5}$.

So what have we accomplished? We started with a quadratic equation and *one* solution for it. We then used polynomial division and computed another solution, and we also saw that these two solutions are the *only ones*!

The point of all this is that this can be done in general, leading us not only to a way of finding one solution of a quadratic equation from another, but, more importantly, a proof of the following

PROPOSITION 5.4. *Any quadratic equation $ax^2+bx+c=0$ (with $a \neq 0$) has at most two real solutions. (In fact, it has at most two complex solutions.)*

The parenthetical statement above can easily be seen once we have covered the complex numbers; so let's concentrate on the case for the real numbers: First of all, we may assume (by dividing both sides by a) that $a=1$, so let's only consider a quadratic equation of the form

$$x^2+px+q=0 \tag{5.38}$$

Furthermore, if this equation has no real solution at all, then we're done. So let's fix one real solution and call it x_0. We will now attempt polynomial division of x^2+px+q by $x-x_0$ and see what happens, as shown in Table 5.3.

Let's go over this very abstract polynomial division in detail: We want to divide x^2+px+q by $x-x_0$. Again, just as in long division, we do so "digit" by "digit", except that instead of "digits", we now use the coefficients: In x^2+px+q , the coefficient 1 of x^2 corresponds to the "hundreds digit", the coefficient p of x corresponds to the "tens

$$
\begin{array}{r|rcrcr}
 & & & x & + & (p+x_0) \\
\hline
x-x_0 & x^2 & + & px & + & q \\
 & -\,(x^2 & - & x_0x) & & \\
\cline{2-4}
 & & & (p+x_0)x & + & q \\
 & & - & ((p+x_0)x & - & (p+x_0)x_0) \\
\cline{4-6}
 & & & & & (p+x_0)x_0+q
\end{array}
$$

TABLE 5.3. Dividing $(x^2+px+q) \div (x-x_0)$ by polynomial division

digit", and the constant coefficient q corresponds to the "ones digit"; in $x-x_0$, the coefficient 1 of x corresponds to the "tens digit", and the constant coefficient $-x_0$ corresponds to the "ones digit".

We first check "how often x goes into" x^2 and find that it "goes x times". So we multiply $x-x_0$ by x and subtract the result from x^2+px+q, leaving us with $(p+x_0)x$. We now "take down" the coefficient q and check again "how often x goes into" $(p+x_0)x$; we find it "goes $(p+x_0)$ times", multiply $x-x_0$ by $p+x_0$, and subtract the result from $(p+x_0)x+q$, leaving us with remainder $(p+x_0)x_0+q$. So x^2+px+q divides into $x-x_0$ with remainder $(p+x_0)x_0+q$, and the quotient is $x+(p+x_0)$. We can write this also as

$$\frac{x^2+px+q}{x-x_0} = (x+(p+x_0)) + \frac{(p+x_0)x_0+q}{x-x_0} \qquad \text{or} \tag{5.39}$$

$$x^2+px+q = (x+(p+x_0))(x-x_0) + ((p+x_0)x_0+q) \tag{5.40}$$

Note that (5.40) is obtained from (5.39) simply by multiplying by $x-x_0$ on both sides. Now let's use the fact that x_0 is a solution of $x^2+px+q=0$: Set $x=x_0$. Then the left-hand side of (5.40) is equal to 0. But the left half of the right-hand side of (5.40) is also equal to 0 for $x=x_0$; so this forces the remainder $(p+x_0)x_0+q$ to be 0 as well, i.e., x^2+px+q divides into $x-x_0$ evenly without remainder! This simplifies (5.40) to

$$x^2+px+q = (x+(p+x_0))(x-x_0)$$

and so, by (5.18), the solutions of $x^2+px+q=0$ are the solutions to the statement

$$x+(p+x_0)=0 \text{ or } x-x_0=0$$

i.e., the solutions are x_0 and $-(p+x_0)$. So there are at most two solutions as claimed. As a by-product, we have also found a formula

which allows us to compute one solution of a quadratic equation from the other. (Note that x_0 may equal $-(p + x_0)$, so there may be only one single solution.)

5.3. A Digression into Square Roots and the Complex Numbers

5.3.1. Square Roots. As you can see from the quadratic formula, solving quadratic equations can necessarily involve square roots, so let's briefly review them first.

By definition, a square root of a real number a is a solution to the equation

$$x^2 = a \tag{5.41}$$

If $a = 0$, then $x = 0$ is a unique solution to this equation. If a is positive, then there are two solutions, namely, a positive number which we denote by $\sqrt{a}$, and its additive inverse, $-\sqrt{a}$. (The existence of such a number $\sqrt{a}$ is actually a very deep fact about the real numbers which we cannot prove here!) In order to make the square root unique for all $a \geq 0$, we define the *non-negative* solution to the equation $x^2 = a$ to be "the" *square root* of a and denote it by $\sqrt{a}$. (Note that this is somewhat arbitrary, it is simply a *definition* of what we mean by "the" square root of a non-negative number. We'll get to what a square root of a negative real number could be later in subsection 5.3.2 when we discuss the complex numbers.)

Let's now state some rules about square roots. The simplest ones are about multiplication and division with square roots: For all non-negative real numbers a and b, we have

$$\sqrt{a} \cdot \sqrt{b} = \sqrt{a \cdot b} \tag{5.42}$$

$$\frac{\sqrt{a}}{\sqrt{b}} = \sqrt{\frac{a}{b}} \tag{5.43}$$

(where we assume $b \neq 0$ for (5.43)). These rules follow immediately from the definition of square roots: Let

$$\begin{aligned}
\sqrt{a} &= \text{ the (unique non-negative) solution } x \text{ to the equation } x^2 = a \\
\sqrt{b} &= \text{ the (unique non-negative) solution } y \text{ to the equation } y^2 = b \\
\sqrt{ab} &= \text{ the (unique non-negative) solution } z \text{ to the equation } z^2 = ab \\
\sqrt{\frac{a}{b}} &= \text{ the (unique non-negative) solution } w \text{ to the equation } w^2 = \frac{a}{b}
\end{aligned}$$

Then (5.42) simply states that

$$xy = z \tag{5.44}$$

and (5.42) states that

$$\frac{x}{y} = w \tag{5.45}$$

Since all the numbers involved are non-negative, we can apply Proposition 5.3: Squaring equations (5.44) and (5.45), we find that they are equivalent to

$$x^2y^2 = (xy)^2 \text{ and } \frac{x^2}{y^2} = \left(\frac{x}{y}\right)^2$$

respectively, both of which are trivially true.

There is one other useful rule about square roots, which connects them to absolute value: For *any* real number a (i.e., also for a negative real number a), we have

$$\sqrt{a^2} = |a| \tag{5.46}$$

This rather surprising (and at first unexpected) fact can be shown directly from the definition of square roots and absolute value, distinguishing the two cases in the definition of absolute value: If $a \geq 0$, then $\sqrt{a^2}$ is the unique non-negative solution x to the equation $x^2 = a^2$ (in the unknown x, with parameter a); but $x = a$ is a non-negative solution to this equation, and since the non-negative solution is unique, we must have $x = a = |a|$. On the other hand, if $a < 0$, then again $\sqrt{a^2}$ is the unique non-negative solution x to the equation $x^2 = a^2$ (in the unknown x, with parameter a!); but $x = -a$ is a non-negative solution to this equation (since $a^2 = (-a)^2$), and since the non-negative solution is unique, we must have $x = -a = |a|$.

We'll conclude this subsection with two "non-rules", which frequently creep into school mathematics. They are the following two statements:

The statement $\sqrt{a+b} = \sqrt{a} + \sqrt{b}$ is generally false!

The statement $\sqrt{a-b} = \sqrt{a} - \sqrt{b}$ is generally false!

Once you realize the above formulas are false, it's very easy to see why: Just plug in almost any numbers and check your answer. Remember: Showing that a "general rule" *fails* is much easier than showing that it *holds*: For a general rule to hold, you must check it in "all" cases, usually by some "mathematical proof"; in order to show that it fails, you need just one counterexample!

5.3.2. The Number i and the Complex Numbers. We now return to the question of how one might define the square root of a negative real number a. Our first attempt would be to define it as a solution x to the equation $x^2 = a$. But there is an obvious problem: for any *real* number x, x^2 is non-negative and thus cannot equal a. So there *cannot* be a real square root of a. But we can extend our number system again.

Think back to the time when you first learned about the negative numbers. Back then, it was impossible to solve an equation like $x+3 = 2$ since that would be the solution to a (non-sensical) problem like

> "Mary receives three apples and now has two apples.
> How many apples did she have before?"

In this context, introducing negative numbers makes no sense: Mary *cannot* have had -1 apples before! But if we expand our model of what numbers mean to notions like temperature, the equation $x + 3 = 2$ suddenly makes sense:

> The temperature rose three degrees and is now at two
> degrees. What was the temperature before?"

We are in a similar situation with solving an equation like $x^2 = -1$. If we insist on staying within the real numbers, there is no solution. The difficulty is that there is no simple and visual model for the complex numbers. The usefulness of the complex numbers can be seen really only in higher mathematics (after most of calculus, in upper-level undergraduate mathematics and beyond) and in fairly sophisticated science and engineering applications. Thus it is far from obvious why one would want to introduce the complex numbers, except for the fact that it enables one to solve some equations.

In order to introduce the complex numbers, let's start with the simplest quadratic equation which cannot be solved in the real numbers, namely, $x^2 = -1$. We introduce a new number i, called the *imaginary unit*, and define it to be a solution of the equation $x^2 = -1$, i.e., we define that $i^2 = -1$. Assuming that all the rules of arithmetic we have for the real numbers also hold for the complex numbers, we see that $(-i)^2 = i^2 = -1$, and so the equation $x^2 = -1$ has two solutions, i and $-i$. Furthermore, we can factor the expression $x^2 + 1$ as $(x + i)(x - i)$.

Let's now move on to solving a slightly more general equation, namely, $x^2 = -a$ for *positive* a. Set $b = \sqrt{a}$. Since we assumed all the rules of arithmetic for the complex numbers, we have

$$(bi)^2 = b^2 i^2 = a(-1) = -a$$

and

$$(-bi)^2 = (-b)^2 i^2 = a(-1) = -a$$

Thus the equation $x^2 = -a$ (for positive a) has the two solutions bi and $-bi$. While one does not in general define one of bi or $-bi$ to be "the" square root of $-a$, both bi and $-bi$ serve as a good substitute for the square root of $-a$. Whenever we write expressions like $\pm\sqrt{-a}$, it doesn't matter anyhow which of bi or $-bi$ we use for $\sqrt{-a}$.

Finally, we'll define a *complex number* to be any number of the form $a + bi$, where a and b are arbitrary real numbers. Addition, subtraction and multiplication are defined on the complex numbers just as expected:

$$(a + bi) + (c + di) = (a + c) + (b + d)i \tag{5.47}$$

$$(a + bi) - (c + di) = (a - c) + (b - d)i \tag{5.48}$$

$$(a + bi)(c + di) = (ac - bd) + (ad + bc)i \tag{5.49}$$

Here, the formula (5.49) for multiplication just uses repeated application of the distributive law for the left-hand side, plus the fact that $i^2 = -1$. Division is somewhat more complicated since one has make the denominator a real number; this involves a slight trick and some messy computations:

$$\begin{aligned} \frac{a + bi}{c + di} &= \frac{(a + bi)(c - di)}{(c + di)(c - di)} \\ &= \frac{(ac + bd) + (bc - ad)i}{c^2 + d^2} \\ &= \frac{ac + bd}{c^2 + d^2} + \frac{bc - ad}{c^2 + d^2} i \end{aligned} \tag{5.50}$$

Note that the last line of (5.50) is well-defined as long as $c+di \neq 0$, i.e., as long as at least one of c and d is nonzero, making $c^2 + d^2$ nonzero. So, for $c + di \neq 0$, the formula (5.50) gives us again an expression of the form $A + Bi$ for real numbers A and B, i.e., a complex number in our sense. (It is also possible to solve quadratic equations with *complex* parameters a, b and c, but that would require the definition of square roots of complex numbers, which is beyond the scope of these notes.)

We now return to the quadratic formula

$$x = -\frac{b}{2a} \pm \frac{\sqrt{b^2 - 4ac}}{2a} \tag{5.36}$$

for the solution to the general quadratic equation

$$ax^2 + bx + c = 0 \tag{5.30}$$

(where a, b and c are now arbitrary real numbers with $a \neq 0$). The quadratic formula then gives us one or two complex solutions in all cases; namely, one solution if $b^2 - 4ac = 0$, and two solutions otherwise.

We conclude with an example: Consider the quadratic equation

$$x^2 + 2x + 2 = 0 \tag{5.51}$$

Here, $a = 1$ and $b = c = 2$. So the solutions to (5.51) are

$$\begin{aligned} x &= -\frac{2}{2} \pm \frac{\sqrt{2^2 - 8}}{2} \\ &= -1 \pm \frac{\sqrt{-4}}{2} \\ &= -1 \pm \frac{2i}{2} \\ &= -1 \pm i \end{aligned}$$

We can check that these two solutions really work:

$$\begin{aligned} (-1+i)^2 + 2(-1+i) + 2 &= (1 - 2i - 1) + (-2 + 2i) + 2 \\ &= (1 - 1 - 2 + 2) + (-2 + 2)i = 0 \end{aligned}$$

and

$$\begin{aligned} (-1-i)^2 + 2(-1-i) + 2 &= (1 + 2i - 1) + (-2 - 2i) + 2 \\ &= (1 - 1 - 2 + 2) + (2 - 2)i = 0 \end{aligned}$$

5.4. Graphing Quadratic Functions and Solving Quadratic Equations Graphically

Solving a quadratic equation $ax^2 + bx + c = 0$ corresponds, of course, to "undoing" the function $f(x) = ax^2 + bx + c$, i.e., to finding all arguments x at which the function f takes the value 0. (We had seen this before in section 2.5 with solving a linear equations $ax + b = 0$, which corresponded to finding all x at which the function $g(x) = ax + b$ takes the value 0.) As in the case of linear equations, one can always find the solutions to quadratic equations by graphing (at least approximately); but the graphs of quadratic functions are naturally a bit more complicated.

Let's start with an example from physics:

> "An arrow is shot straight up from the ground with an initial velocity of $v_0 = 50$m/sec. If the height of the arrow (in meters) at time t (after being fired, in seconds) is described by

$$s = v_0 t - \frac{1}{2} g t^2 \tag{5.52}$$

(neglecting friction, where $g \approx 9.8\text{m/sec}^2$ is the earth's gravity constant), how high will the arrow go, and when will it hit the ground again?"

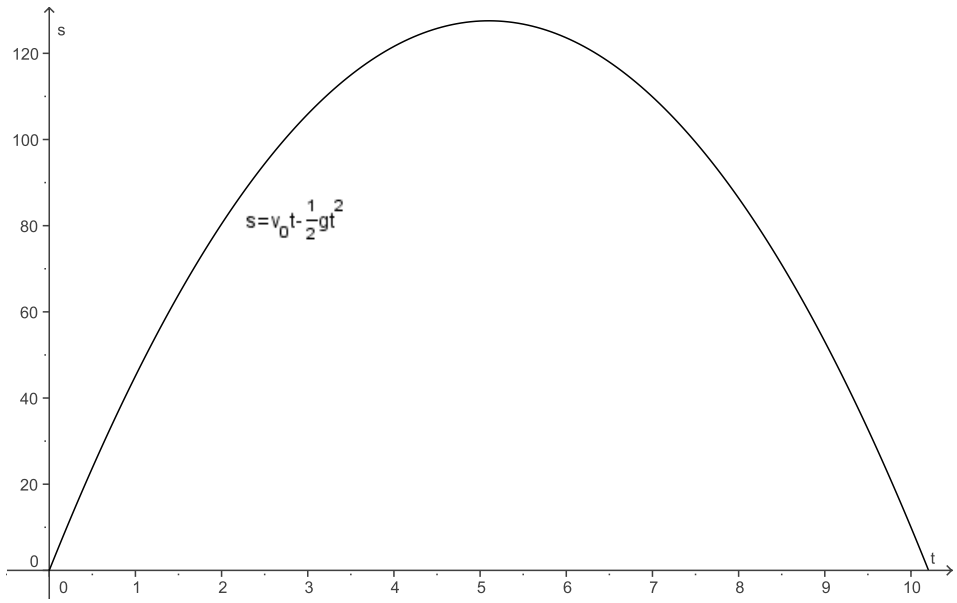

FIGURE 5.7. The motion of an arrow

The motion of the arrow is described approximately by the graph of the function

$$s(t) = 50t - 4.9t^2 \tag{5.53}$$

shown in Figure 5.7. (We only show the graph for values of t in the interval $[0, 10.2]$ since the graph appears to be below the t-axis outside this interval, which would correspond to the arrow being below the ground, contradicting the "real life" context.) It is now not hard to check that the arrow is at maximum height at just over five seconds, when it reaches a height of approximately 130 meters. Furthermore, the arrow appears to hit the ground after just over ten seconds.

Could we have computed this without a picture of the graph? Of course, but that requires a bit of thinking and computation. Let's address the second question from the problem first: When does the arrow hit the ground again? This corresponds to a solution of the quadratic equation

$$50t - 4.9t^2 = 0$$

since the arrow hits the ground when s equals 0 again. We can solve this equation by factoring:

$$-4.9t^2 + 50t = -4.9t(t - \frac{50}{4.9}) = 0$$

and so s equals 0 at time $t = 0$ (when the arrow is fired) and at time $t \approx \frac{50}{4.9} \approx 10.2$. The latter time (10.2 seconds after being fired) must be the approximate time when the arrow hits the ground again.

The first question (How high will the arrow go, and when will it be at maximum height?) is harder to answer.[7] The solution uses our technique of completing the square from subsection 5.2.4:

$$\begin{aligned} s &= 50t - 4.9t^2 \\ &= -4.9\left(t^2 - \tfrac{50}{4.9}t\right) \\ &= -4.9\left(t^2 - \tfrac{50}{4.9}t + \left(\tfrac{25}{4.9}\right)^2 - \left(\tfrac{25}{4.9}\right)^2\right) \\ &= -4.9\left(t^2 - \tfrac{50}{4.9}t + \left(\tfrac{25}{4.9}\right)^2\right) + \tfrac{25^2}{4.9} \\ &= -4.9\left(t - \tfrac{25}{4.9}\right)^2 + \tfrac{25^2}{4.9} \end{aligned} \tag{5.54}$$

Since $-4.9\left(t - \frac{25}{4.9}\right)^2 \leq 0$ for all t, we can now see that the height s cannot exceed $\frac{25^2}{4.9} \approx 127.5$ meters, but that that height is indeed the maximum height and is achieved at $\frac{25}{4.9} \approx 5.1$ seconds.

Now let's do the above in the abstract. Consider a *quadratic function*

$$f(x) = ax^2 + bx + c \tag{5.55}$$

where $a \neq 0$. We can rewrite this, completing the square, as

$$\begin{aligned} f(x) &= a\left(x^2 + \frac{b}{a}x\right) + c \\ &= a\left(x^2 + \frac{b}{a}x + \left(\frac{b}{2a}\right)^2 - \left(\frac{b}{2a}\right)^2\right) + c \\ &= a\left(x^2 + \frac{b}{a}x + \left(\frac{b}{2a}\right)^2\right) + c - \frac{b^2}{4a} \\ &= a\left(x + \frac{b}{2a}\right)^2 + \left(c - \frac{b^2}{4a}\right) \text{ and so} \end{aligned} \tag{5.56}$$

[7] Calculus was invented to solve problems like this, but in our easy case, we don't need it!

and so

$$f(x) - \left(c - \frac{b^2}{4a}\right) = a\left(x - \left(-\frac{b}{2a}\right)\right)^2 \tag{5.57}$$

Setting

$$x_0 = \left(-\frac{b}{2a}\right) \text{ and } y_0 = \left(c - \frac{b^2}{4a}\right)$$

equation (5.57) simplifies to

$$f(x) - y_0 = a(x - x_0)^2 \qquad \text{or} \qquad f(x) = a(x - x_0)^2 + y_0 \tag{5.58}$$

The graph of the quadratic function is called a *parabola*, and the point V with coordinates (x_0, y_0) is called the *vertex* of the parabola. (In the arrow example (5.53) above, the vertex is approximately the point $(5.1, 127.5)$.)

We are still assuming that $a \neq 0$ and now need to distinguish two cases before we can discuss the graph of a quadratic function in more detail:

Case 1: $a > 0$: Then $a(x - x_0)^2 + y_0 \geq y_0$ for all x, and so the minimal value of f is y_0, which is achieved only at $x = x_0$. The equation describes a parabola which "opens upward". In order to solve the quadratic equation

$$a(x - x_0)^2 + y_0 = 0 \tag{5.59}$$

we need to distinguish three subcases:

Case 1a: $y_0 < 0$: Then the equation (5.59) has two real solutions

$$x = x_0 \pm \sqrt{-\frac{y_0}{a}}$$

since the graph of the function $f(x) = a(x - x_0)^2 + y_0$ intersects the x-axis at the two points $(x_0 + \sqrt{-\frac{y_0}{a}}, 0)$ and $(x_0 - \sqrt{-\frac{y_0}{a}}, 0)$. The function achieves its minimum value at the vertex (x_0, y_0). (See Figure 5.8 for an example with $x_0 = 2$, $y_0 = -1$, and so the vertex is $V = (2, -1)$; the graph of f intersects the x-axis at $(1, 0)$ and $(3, 0)$.)

Case 1b: $y_0 = 0$: Then the equation (5.59) has only one real solution

$$x = x_0$$

since the graph of the function $f(x) = a(x - x_0)^2 + y_0$ intersects the x-axis only at the vertex $(x_0, 0)$, at which the function also achieves its minimum value. (See Figure 5.9 for an example with $x_0 = 2$, $y_0 = 0$, so the vertex

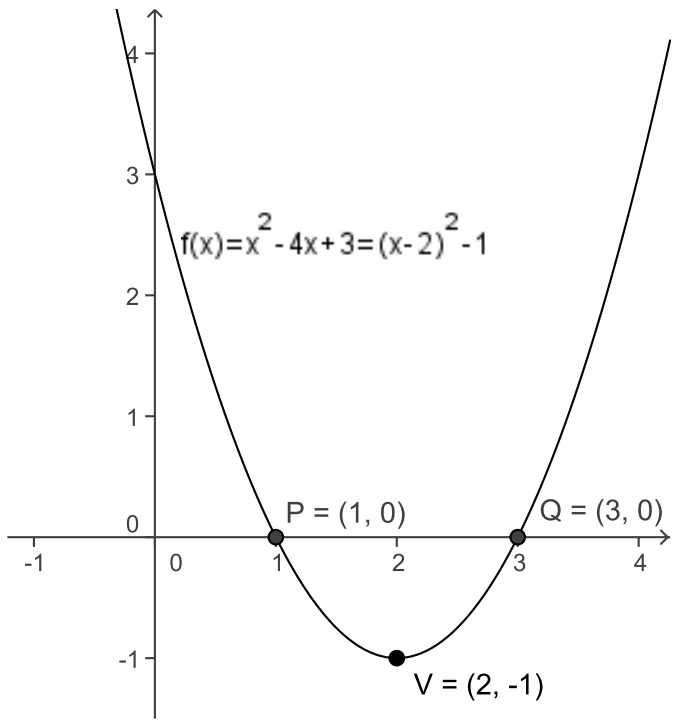

FIGURE 5.8. The graph of the quadratic function $f(x) = x^2 - 4x + 3$

is $V = (2, 0)$, which is also the point where the graph of f intersects the x-axis.)

Case 1c: $y_0 > 0$: Then the equation (5.59) has no real solutions since

$$f(x) = a(x - x_0)^2 + y_0 \geq y_0 > 0$$

for all t. The function still achieves its minimum value at the vertex (x_0, y_0). (See Figure 5.10 for an example with $x_0 = 2$, $y_0 = 1$, so the vertex is $V = (2, 1)$, and the graph of f does not intersect the x-axis.)

Case 2: $a < 0$: Then $a(x - x_0)^2 + y_0 \leq y_0$ for all x, and so the maximal value of f is y_0, which is achieved only at $x = x_0$. The equation describes a parabola which "opens downward". In order to solve the quadratic equation

$$a(x - x_0)^2 + y_0 = 0 \tag{5.59}$$

we again need to distinguish three subcases:

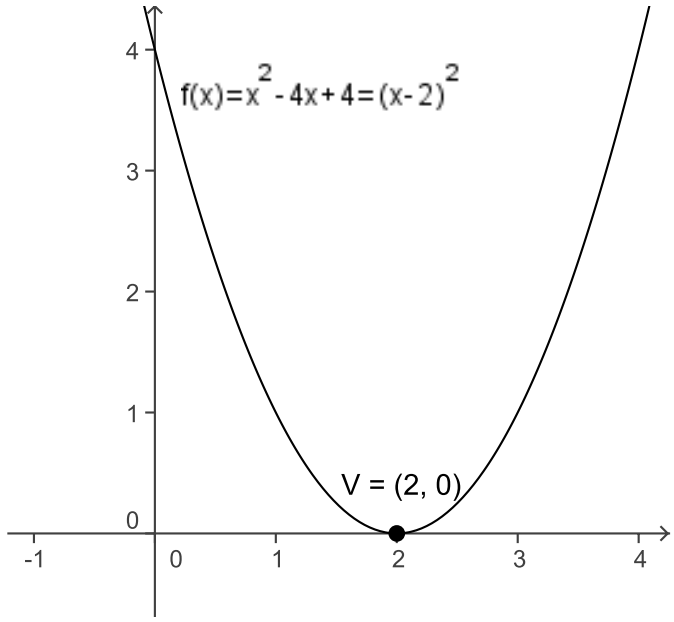

FIGURE 5.9. The graph of the quadratic function $f(x) = x^2 - 4x + 4$

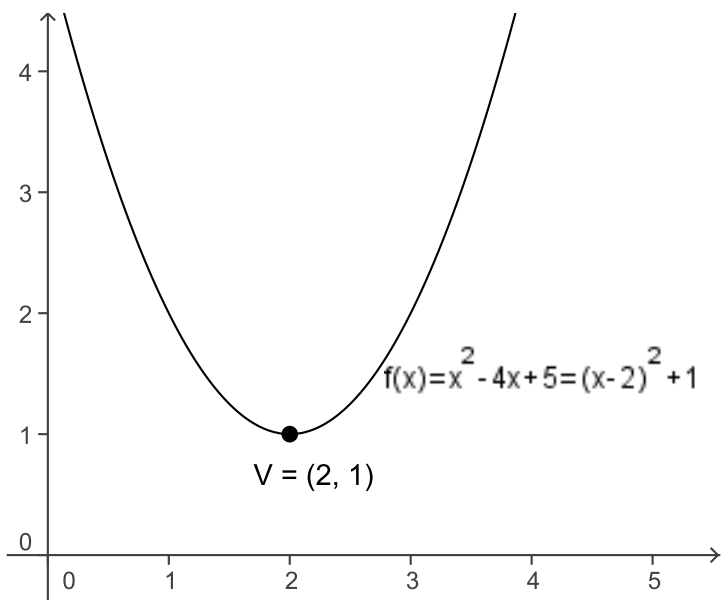

FIGURE 5.10. The graph of the quadratic function $f(x) = x^2 - 4x + 5$

Case 2a: $y_0 > 0$: Then the equation (5.59) has two real solutions

$$x = x_0 \pm \sqrt{-\frac{y_0}{a}}$$

since the graph of the function $f(x) = a(x - x_0)^2 + y_0$ intersects the x-axis at the two points $(x_0 + \sqrt{-\frac{y_0}{a}}, 0)$ and $(x_0 - \sqrt{-\frac{y_0}{a}}, 0)$. The function achieves its maximum value at the vertex (x_0, y_0). (See Figure 5.11 for an example with $x_0 = -2$, $y_0 = 1$, and so the vertex is $V = (-2, 1)$; the graph of f intersects the x-axis at $(1, 0)$ and $(3, 0)$.)

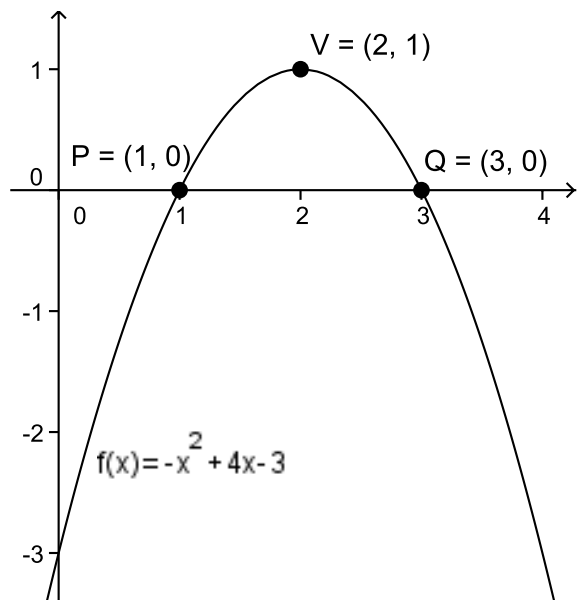

FIGURE 5.11. The graph of the quadratic function $f(x) = -x^2 + 4x - 3$

Case 2b: $y_0 = 0$: Then the equation (5.59) has only one real solution

$$x = x_0$$

since the graph of the function $f(x) = a(x - x_0)^2 + y_0$ intersects the x-axis only at the vertex $(x_0, 0)$, at which it also achieves its maximum value.

Case 2c: $y_0 < 0$: Then the equation (5.59) has no real solutions since

$$f(x) = a(x - x_0)^2 + y_0 \leq y_0 < 0$$

for all t. The function still achieves its maximum value at the vertex (x_0, y_0).

Finally, let's study the effect of varying the three parameters a, x_0 and y_0 in the equations of a quadratic function

$$f(x) - y_0 = a(x - x_0)^2 \qquad \text{or} \qquad f(x) = a(x - x_0)^2 + y_0 \tag{5.58}$$

Since the vertex of the parabola described by these equations is always the point (x_0, y_0), the effect of changing x_0 or y_0 is clear: It shifts the whole parabola right or left (when changing x_0) or up or down (when changing y_0), respectively. The effect of changing a is a bit more subtle. We saw already that if $a > 0$ then the parabola "opens upward", and that it if $a < 0$ then the parabola "opens downward". The effect of increasing or decreasing $|a|$ is now to make the parabola "steeper" or less "steep". (Compare Figure 5.8 where $a = 1$ with Figure 5.12 where $a = 4$.)

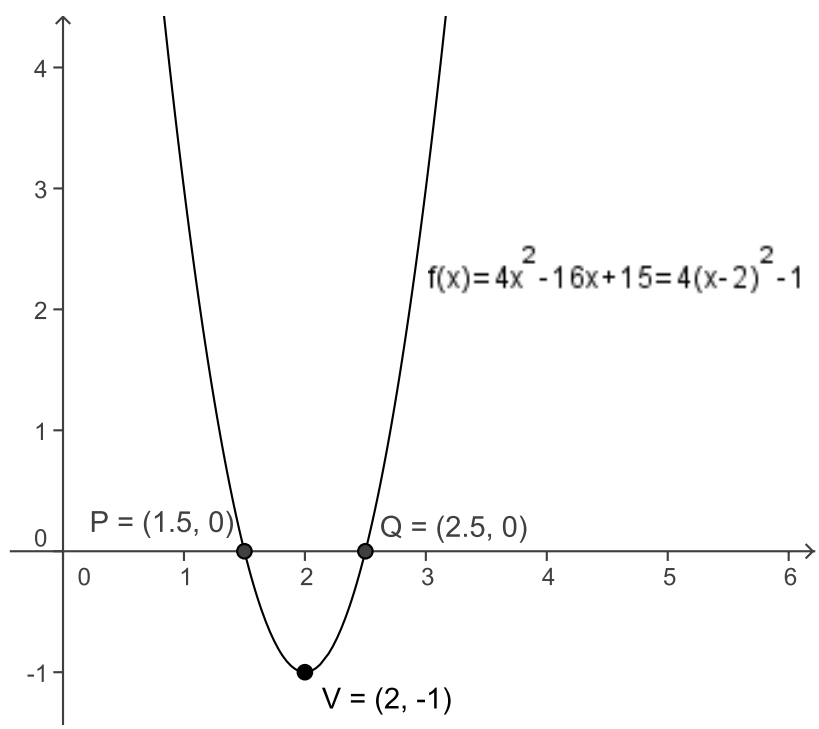

FIGURE 5.12. The graph of the quadratic function $f(x) = 4x^2 - 16x + 15$

5.5. Quadratic Inequalities

We conclude this chapter by considering quadratic inequalities. Let's vary the arrow problem from section 5.4 a bit:

> "An arrow is shot straight up from the ground with an initial velocity of $v_0 = 50$m/sec. If the height of the arrow (in meters) at time t (after being fired, in seconds) is described by
>
> $$s = v_0 t - \frac{1}{2} g t^2 \tag{5.52}$$
>
> (neglecting friction, where $g \approx 9.8\text{m/sec}^2$ is the earth's gravity constant), when will the arrow be at least 50 meters above the ground?"

We already saw that the motion of the arrow is described approximately by the graph of the function

$$s(t) = 50t - 4.9t^2 \tag{5.53}$$

shown in Figure 5.7. So answering the above problem amounts to solving the *quadratic inequality*

$$50t - 4.9t^2 \geq 50 \tag{5.60}$$

First note a common mistake: We cannot simply rewrite this inequality as $-4.9t^2 + 50t - 50 \geq 0$ and "use the quadratic formula" to get $t \geq \dfrac{-50 \pm \sqrt{50^2 - 4 \cdot (-4.9) \cdot (-50)}}{2 \cdot (-4.9)}$; this "solution" doesn't even make any sense since it specifies t to be $\geq$ some number $\pm$ some other number! Instead, we can either use the graph of the function or use the technique of completing the square.

For a graphical solution, we simply look at the graph in Figure 5.7 and note that the graph is at or above the line $s = 50$ approximately between times $t = 1\frac{1}{4}$ seconds and $t = 9$ seconds.

For an exact solution, we use the "completed square version" of equation (5.53):

$$s = -4.9\left(t - \tfrac{25}{4.9}\right)^2 + \tfrac{25^2}{4.9} \tag{5.54}$$

We can now solve our inequality in a few steps as follows:

$$-4.9\left(t-\tfrac{25}{4.9}\right)^2+\tfrac{25^2}{4.9}\geq 50 \tag{5.61}$$

$$-4.9\left(t-\tfrac{25}{4.9}\right)^2\geq 50-\tfrac{25^2}{4.9} \tag{5.62}$$

$$\left(t-\frac{25}{4.9}\right)^2\leq\frac{50-\frac{25^2}{4.9}}{-4.9}=\frac{\frac{25^2}{4.9}-50}{4.9}=\frac{25^2-4.9\cdot 50}{4.9^2} \tag{5.63}$$

$$\left|t-\frac{25}{4.9}\right|\leq\sqrt{\frac{25^2-4.9\cdot 50}{4.9^2}} \tag{5.64}$$

$$\frac{25}{4.9}-\sqrt{\frac{25^2-4.9\cdot 50}{4.9^2}}\leq t\leq\frac{25}{4.9}+\sqrt{\frac{25^2-4.9\cdot 50}{4.9^2}} \tag{5.65}$$

which gives the approximate solution

$$1.124\leq t\leq 9.08$$

Looking over the steps in (5.61)–(5.65), we see that the key steps are in going from (5.63) to (5.64) and from (5.64) to (5.65). The second of these two steps involves statements (4.15) and (4.16) about the absolute value, namely, combining the two, that we have

$$|a|\leq r \text{ exactly when } -r\leq a\leq r \tag{5.66}$$

The first of these two steps, going from (5.63) to (5.64), involves a modification of Proposition 5.3 to the case of inequalities:

PROPOSITION 5.5. *Let A be an algebraic expression and $r\geq 0$ a number. Then:*

(1) The solutions to the equation

$$A^2=r$$

are the same as the solutions to the statement

$$A=\sqrt{r} \text{ or } A=-\sqrt{r}$$

(2) The solutions to the inequality

$$A^2<r$$

are the same as the solutions to the statement

$$-\sqrt{r}<A<\sqrt{r}$$

(3) The solutions to the equation

$$A^2>r$$

are the same as the solutions to the statement

$$A<-\sqrt{r} \text{ or } A>\sqrt{r}$$

Proposition follow immediately from (4.15)–(4.17) and from the fact that for all *non-negative* numbers a and b:

$$a < b \text{ exactly when } a^2 < b^2 \tag{5.67}$$

We now see that the key step above, going from (5.63) to (5.64), follows immediately from Proposition 5.5.

It is also not hard to see that we could now carry out a general analysis and solution for quadratic inequalities of the form $ax^2 + bx + c \leq 0$, < 0, ≥ 0 or > 0, respectively, in the style of section 5.2 (for an algebraic solution) or section 5.4 (for a graphical solution).

CHAPTER 6

Exponential and Logarithmic Functions, Equations and Inequalities

In this last chapter, we'll cover the exponential functions and their inverse functions, the logarithmic functions.

6.1. Exponentiation: A Review of the Definition

There is one common arithmetical operation from school mathematics which we have mostly ignored so far: exponentiation, or "taking the power". This operation is initially introduced as repeated multiplication, just as multiplication (in the whole numbers) can be viewed as repeated addition.

The actual definition of exponentiation is quite tricky, and we will present it here very carefully, in a number of steps.

Given any number a and any nonzero whole number m, we define the *mth power of* a to be

$$a^m = \overbrace{a \cdot a \cdot a \ldots a}^{m} \tag{6.1}$$

Here (and from now on), we will call a the *base* and m the *exponent* (or the *power*)[1]. We read a^m as "a to the mth power" (or "a to the mth" for short). For $m = 2$, we read a^2 as "a squared" (or "the square of a"); and for $m = 3$, we read a^3 as "a cubed" (or "the cube of a").

This definition of exponentiation as repeated multiplication gives us immediately the following rules of exponentiation (for now, only for nonzero whole numbers m and n):

$$a^m \cdot a^n = a^{m+n} \tag{6.2}$$

$$\frac{a^m}{a^n} = a^{m-n} \text{ (if } m > n) \tag{6.3}$$

$$(a^m)^n = a^{m \cdot n} \tag{6.4}$$

$$(a \cdot b)^m = a^m \cdot b^m \tag{6.5}$$

[1]In the Singapore math books, the exponent is also called the *index* (plural: indices).

To see that, e.g., (6.2) is true, simply observe that

$$\begin{aligned} a^m \cdot a^n &= \overbrace{a \cdot a \cdot a \cdots a}^{m} \cdot \overbrace{a \cdot a \cdot a \cdots a}^{n} \\ &= \overbrace{a \cdot a \cdot a \cdots a}^{m+n} \\ &= a^{m+n} \end{aligned}$$

And to see that (6.4) is true, we check

$$\begin{aligned} (a^m)^n &= \overbrace{\overbrace{a \cdot a \cdot a \cdots a}^{m} \cdot \overbrace{a \cdot a \cdot a \cdots a}^{m} \cdot \overbrace{a \cdot a \cdot a \cdots a}^{m} \ldots \overbrace{a \cdot a \cdot a \cdots a}^{m}}^{n} \\ &= \overbrace{a \cdot a \cdot a \cdots a}^{m \cdot n} \\ &= a^{m \cdot n} \end{aligned}$$

Exercise 6.1. How you would prove (6.3) and (6.5) this way? Why does it make sense to impose the restriction $m > n$ in (6.3) for now?)

So (6.2)–(6.5) follow immediately from our previous rules for multiplication. Wouldn't it be nice to have these rules hold not just for nonzero whole numbers as exponents? Of course, you know that these rules are still true, right?

Well, there is a small problem: We haven't even *defined* what a^m *means* unless m is a nonzero whole number! We will do this in a sequence of small steps, each time justifying our definition for more and more possible exponents.

From now on, we will assume that the base a is nonzero. (This will make sense in a moment when we see how division by a becomes involved in our definition.) Now we can define exponentiation for an exponent which is 0 or a negative integer as follows:

$$a^0 = 1 \tag{6.6}$$

$$a^m = \frac{1}{a^{-m}} \text{ (for any negative integer } m) \tag{6.7}$$

First, note here that a^{-m} in (6.7) has already been defined in (6.1) since $-m$ is a nonzero whole number if m is a negative integer. More importantly, note that we are *forced* to define exponentiation this way if we want the analogue of (6.3) to still hold:

$$\begin{aligned} a^0 &= \frac{a^1}{a^1} = \frac{a}{a} = 1 \\ a^m &= \frac{a^0}{a^{-m}} = \frac{1}{a^{-m}} \end{aligned}$$

What about rational exponents? For this, we need to restrict the base a even further: We'll assume from now on that a is not only nonzero but indeed a positive real.[2] Now we'll define exponentiation for rational exponents as follows (for any integer n and any positive integer p):

$$a^{\frac{1}{p}} = \sqrt[p]{a} \tag{6.8}$$

$$a^{\frac{n}{p}} = \left(a^{\frac{1}{p}}\right)^n \tag{6.9}$$

Of course, this immediately raises another problem: We haven't defined yet what $\sqrt[p]{a}$ means for a positive real number a and a positive integer p. But (as for the case $p = 2$, i.e., the square root), we can define the *pth root* of a simply as the unique positive solution[3] to the equation $x^p = a$. Again note also that the definition in (6.9) already requires the definition in (6.8). And again, more importantly, note that we are *forced* to define exponentiation this way if we want the analogue of (6.4) to still hold:

$$\begin{aligned} \left(a^{\frac{1}{p}}\right)^p &= a^{\frac{p}{p}} = a^1 = a \\ a^{\frac{n}{p}} &= a^{\frac{1}{p}\cdot n} = \left(a^{\frac{1}{p}}\right)^n \end{aligned}$$

But there are two hidden problems with these definitions:

(1) It is possible that the rational number $\frac{n}{p}$ (for an integer n and a positive integer p) happens to equal an integer m, say. Now we need to show that the definition in (6.9) coincides with the definition from (6.1) and (6.6)–(6.7). In that case, we have $m = \frac{n}{p}$ and so $n = mp$. Now we can argue as follows

[2]We impose this restriction since we'll be taking roots in a moment. It is actually possible to define a^m in a meaningful way for a negative real and *some* rational exponents m, but the rules for when this is possible are rather complicated, so we'll ignore this case here.

[3]The fact that such a solution always exists and is unique is hard to show and beyond the scope of these notes; so we'll simply assume it. For an *odd* positive integer p, one can also define $\sqrt[p]{a}$ for negative a this way; but we'll not do so here.

(using (6.4) for integer exponents only!):

$$\begin{aligned} a^{\frac{n}{p}} &= \left(a^{\frac{1}{p}}\right)^n \\ &= \left(a^{\frac{1}{p}}\right)^{p\cdot m} \\ &= \left(\left(a^{\frac{1}{p}}\right)^p\right)^m \\ &= a^m \end{aligned}$$

(2) Similarly, given a rational number r, it can be written in many different ways as a fraction. So, in order to show that a^r is uniquely defined, we really need to show that whenever $\frac{n}{p} = \frac{n'}{p'}$ for integers n and n' and for positive integers p and p', then

$$a^{\frac{n}{p}} = a^{\frac{n'}{p'}}$$

under the definition in (6.9). This proof is similar to the one in (1) but quite a bit more tedious since it involves many cases, so we'll skip it here.

Finally, can we define a^r for *irrational* exponent r? The answer is a qualified "yes"; qualified because in order to make this definition precise, one really needs calculus, which we clearly don't want to assume here. Instead, let's give the intuition.[4]

For an irrational number r and arbitrary positive real number a, we define a^r as being more and more closely approximated by a^q where q is a rational number close to r. The fact that this definition works is a "miracle" we have to accept here, since proving it would take us far beyond the scope of these notes.

Finally, looking back, we really need to return to checking our rules of exponentiation (6.2)–(6.5) for arbitrary exponents. We state them again here for reference as a proposition:

[4]Strictly speaking, what we are doing here is merely what you have already done for the other arithmetical operations, like addition and multiplication, for arbitrary real numbers. E.g., the sum $r_0 + r_1$ of two irrational numbers r_0 and r_1, say, is "approximated" more and more "closely" by the sum $q_0 + q_1$ of rational numbers q_0 and q_1 "close" to r_0 and r_1, respectively.

PROPOSITION 6.2. *The following* exponential rules *hold for all positive real numbers a and b, and for all real numbers r and s:*

$$a^r \cdot a^s = a^{r+s} \tag{6.10}$$

$$\frac{a^r}{a^s} = a^{r-s} \tag{6.11}$$

$$(a^r)^s = a^{r \cdot s} \tag{6.12}$$

$$(a \cdot b)^r = a^r \cdot b^r \tag{6.13}$$

If both r and s are integers, then these rules hold for arbitrary nonzero real numbers a and b.

The formal proof of Proposition 6.2 is quite tedious, so we'll skip it here and simply accept the exponential rules as given from now on.

There is one special case of (6.11) which comes up frequently, namely, the case when $r = 0$, so we'll state this as a separate formula:

$$a^{-s} = \frac{1}{a^s} = \left(\frac{1}{a}\right)^s \quad \text{or, equivalently:} \quad a^t = \frac{1}{a^{-t}} = \left(\frac{1}{a}\right)^{-t} \tag{6.14}$$

6.2. Motivation for Exponentiation

So far, we have only discussed the definition of exponentiation in fair detail and have left the motivation rather weak: "Exponentiation is repeated multiplication, just like multiplication is repeated addition." Of course, this motivation is not really satisfying; one could now go on and talk about "repeated exponentiation", but there is a really good reason why almost no one does: There are essentially no real-life phenomena which can be described by "repeated exponentiation", but there are quite a few which can be described by exponentiation.

Let's look at the following two examples with which I'd like to contrast linear vs. exponential functions:

> "A fish in Ann's aquarium is giving birth to babies. Every hour, three new babies are born. If she has five fish at the beginning, how many fish will she have after n hours?"
>
> "The number of bacteria in a lab dish doubles every hour. If there are one hundred bacteria at the beginning, how many will there be after n hours?"

Look at the first problem first; Table 6.1 shows the number of fish during the first few hours.

We see right away that the number of fish increases each hour by a fixed amount, namely, by 3. Therefore the number F of fish after n

number of hours	0	1	2	3	4	5
number of fish	5	8	11	14	17	20

TABLE 6.1. The number of fish after n hours

hours is given by the linear function

$$F(n) = 3n + 5$$

or, more awkwardly, for reasons which will become apparent in a few paragraphs, as

$$F(n) = 5 + 3 \cdot n \tag{6.15}$$

(Note that this function really only makes sense for whole number values of n; it doesn't make sense to talk about there being 9.5 fish after 1.5 hours! But let's ignore this side issue for now.)

Now let's look at the second problem; Table 6.2 shows the number of bacteria during the first few hours.

number of hours	0	1	2	3	4	5
number of bacteria	100	200	400	800	1,600	3,200

TABLE 6.2. The number of bacteria after n hours

We see right away that the number of bacteria does *not* increase by a fixed amount each hour. In fact, the increase itself increases rather dramatically: From hour 0 to hour 1, the increase is by only 100, but from hour 4 to hour 5, the increase is already by 1,600. In fact, this leads us to the following observation: If we denote the number of bacteria after n hours by $B(n)$, and consider the "first difference function" (as defined in section 5.1), namely,

$$B^1(n) = B(n+1) - B(n)$$

then it turns out that in fact $B^1(n) = B(n)$, i.e., the first difference function B^1 equals the original function B. We could now try, as in section 5.1, to take the second difference function

$$B^2(n) = B^1(n+1) - B^1(n)$$

but, of course, this will give us the same function again! So not only is the function B not linear; it is also not quadratic, by the discussion in section 5.1.

On the other hand, you will surely have noticed definite "patterns" in Tables 6.1 and 6.2: In the former, the *additive change* is constant; namely, each hour, the number of fish increases by a *summand* of 3. On the other hand, in the latter, the *multiplicative change* is constant; namely, each hour, the number of bacteria increases by a *factor* of 2.

It is then not hard to see that the function B can be described by

$$B(n) = 100 \cdot 2^n \tag{6.16}$$

Note that there is some similarity between (6.15) and (6.16): The "starting point" (at $n = 0$) is 5 and 100, respectively, and shows up first in the expression on the right-hand side. Next comes the operation symbol for the way in which the increase occurs, additively in (6.15), and multiplicatively in (6.16). Next comes the "rate of change", namely, 3 in (6.15) for the additive change of 3 each hour, and 2 in (6.16) for the multiplicative change of 2 each hour, respectively. And finally comes n for the number of times by which this increase occurs, preceded by symbol for the "repeated operation" (namely, multiplication as repeated addition in (6.15), and exponentiation as repeated multiplication in (6.16), respectively).

In both our fish and our bacteria example, it is now easy to compute the number of fish or bacteria, respectively, at time 10 hours, say: It is $5 + 3 \cdot 10 = 35$ fish and $100 \cdot 2^{10} = 102,400$ bacteria, respectively.[5] But, as we mentioned already in section 2.5, such problems are really problems of arithmetic: In order to solve them, we merely have to evaluate the numerical expressions $5 + 3 \cdot 10$ and $100 \cdot 2^{10}$, respectively.

On the other hand, suppose that we are given the number of fish, or bacteria, respectively, and want to compute the number of hours after which there are this many fish or bacteria. Suppose I want to know when there will be 50 fish. This is now a "real algebra" problem and clearly amounts to solving the equation

$$5 + 3 \cdot n = 50$$

by "undoing" the algebraic expression $5 + 3 \cdot n$ to arrive, step by step, at the solution

$$\begin{aligned} 3n &= 50 - 5 \\ n &= \frac{50 - 5}{3} = 15 \end{aligned}$$

[5]In the latter case, of course, we should really write "approximately 100,000 bacteria" since 102,400 gives a false sense of precision for our answer.

So it will be after 15 hours that there will be 50 fish in the aquarium; and in order to arrive at the solution, we first subtract 5 from both sides of the equation, and then divide by 3 on both sides.

Let's try the same for the bacteria: Suppose I want to know after how many hours there will be 6,400 bacteria? Again, this is now a "real algebra" problem and clearly amounts to solving the equation

$$100 \cdot 2^n = 6,400$$

by "undoing" the algebraic expression $100 \cdot 2^n$ to arrive, step by step, at the solution

$$\begin{aligned} 2^n &= \frac{6,400}{100} = 64 \\ n &= 6 \end{aligned}$$

So it will be after 6 hours that there will be 6,400 bacteria in the lab dish. However, obtaining this solution is conceptually more difficult: In the first step, we simply divide both sides by 100. But the second step, going from $2^n = 64$ to $n = 6$, is a bit mysterious at this point. We will learn how to solve such so-called "exponential equations" in general in section 6.4, when we introduce the logarithmic functions. In our particular example with small numbers, we can, of course, just use "guess-and-check".[6]

6.3. The Exponential Functions

6.3.1. Properties of Exponential Functions. Before we proceed to solving exponential equations (such as $2^n = 64$ in the example above), let's talk about exponential functions in general and note some of their properties.

First of all, for each possible *base* $b > 0$, we can define an *exponential function* $f(x) = b^x$, which is defined for all real numbers x.[7] In the case that $b = 1$, we get the rather boring constant function $f(x) = 1^x = 1$, so we will also assume $b \neq 1$ for the remainder of this section.

We first look at two typical examples of exponential functions, the function $f(x) = 2^x$ (see Figure 6.1 for its graph), and the exponential function $g(x) = \left(\frac{1}{2}\right)^x$ (see Figure 6.2 for its graph).

From these two graphs, we observe some properties which we state in the following proposition. We will then argue that all exponential functions "look like one these two functions", depending on whether $b > 1$ or $0 < b < 1$, respectively.

[6] And I had to start with the rather artificial number of 6,400 bacteria to "make the problem work out".

[7] We explained already in section 6.1 why the assumption $b > 0$ is needed.

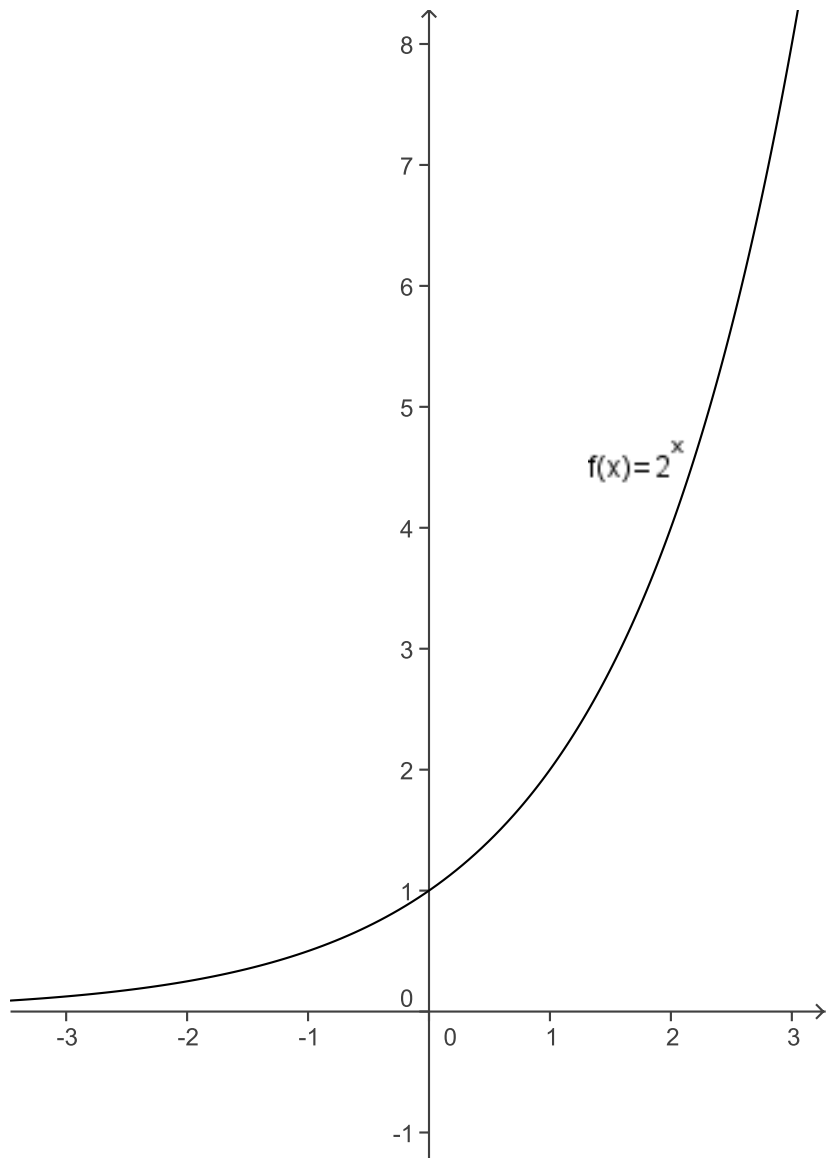

FIGURE 6.1. The graph of the exponential function $f(x) = 2^x$

PROPOSITION 6.3. *For any base $b > 0$ (with $b \neq 1$), the exponential function $h(x) = b^x$ has the following properties:*

(1) The exponential function $h(x) = b^x$ is defined for all real numbers x, i.e., the domain of h is the set $\mathbf{R}$ of all real numbers.
(2) The graph crosses the y-axis at the point $(1, 0)$.
(3) The graph lies entirely above the x-axis, i.e., $b^x > 0$ for all x.

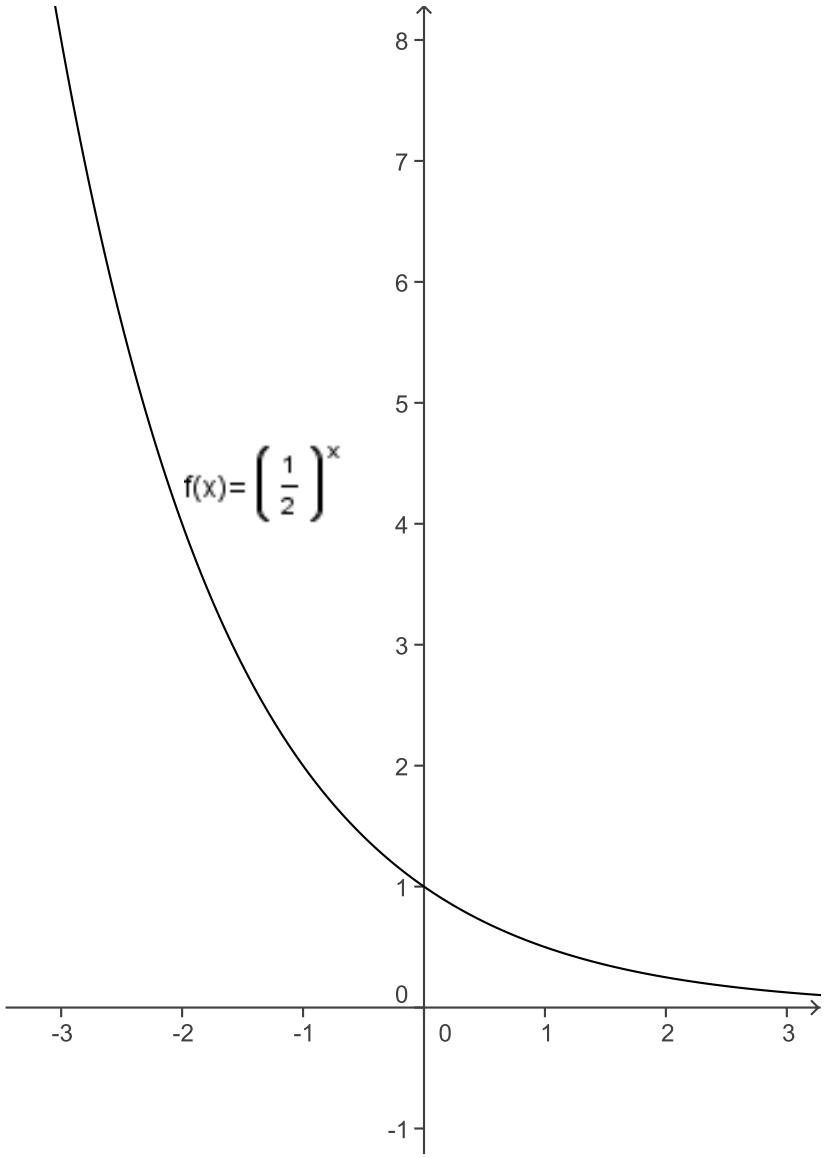

FIGURE 6.2. The graph of the exponential function $f(x) = \left(\frac{1}{2}\right)^x$

(4) If $b > 1$ then

- *(a) the exponential function $h(x) = b^x$ increases as x increases, i.e., if $x_0 < x_1$ then $b^{x_0} < b^{x_1}$;*
- *(b) the graph increases "without bound" as x grows larger and larger; and*

(c) the graph approaches the x-axis arbitrarily closely as x grows smaller and smaller (i.e., as x "grows larger and larger in the negative direction").

(5) If $0 < b < 1$ then

(a) the exponential function $h(x) = b^x$ decreases as x increases, i.e., if $x_0 < x_1$ then $b^{x_0} > b^{x_1}$;

(b) the graph approaches the x-axis arbitrarily closely as x grows larger and larger; and

(c) the graph increases "without bound" as x grows smaller and smaller (i.e., as x "grows larger and larger in the negative direction").

(6) The exponential function $h(x) = b^x$ is 1–1.

(7) The range of the exponential function $h(x) = b^x$ is the set $(0, \infty)$ of all positive real numbers.

In summary, the exponential function $h(x) = b^x$ is a 1–1 function with domain $\mathbf{R}$ and range $(0, \infty)$ (the set of all positive real numbers).

Let's look at these statements one by one and try to argue that they hold true not only for the special cases $b = 2$ and $b = \frac{1}{2}$ shown in Figures 6.1 and 6.2, but indeed for all positive bases $b \neq 1$.

First of all, (1)–(3) follow by analyzing the definitions in section 6.1.

Assume that $b > 1$. Then, for (4a), note that $x_0 < x_1$ implies by (6.11) that

$$\frac{b^{x_1}}{b^{x_0}} = b^{x_1 - x_0} > 1$$

since $x_1 - x_0 > 0$, and so $b^{x_1} > b^{x_0}$. Next, (4b) and (4c) are somewhat harder to check: In the special case $b = 2$, we see that $x < 2^x$ for all x; more generally, for arbitrary $b > 1$, we have that $(b-1)x < b^x$. But now $(b-1)x$ grows "without bound" as x does, and so must b^x. Finally, (4c) follows by (6.14), since, as x grows smaller and smaller (i.e., as x "grows larger and larger in the negative direction"), $-x$ grows larger and larger ("in the positive direction"), but then, by (6.14),

$$b^x = \frac{1}{b^{-x}}$$

and this expression gets closer and closer to 0 since b^{-x} grows without bound.

Now assume that $0 < b < 1$. Then, similarly, for (5a), note that $x_0 < x_1$ implies by (6.11) that

$$\frac{b^{x_1}}{b^{x_0}} = b^{x_1 - x_0} < 1$$

since $x_1 - x_0 > 0$, and so $b^{x_1} < b^{x_0}$. Next, (5b) and (5c) can be reduced to (4c) and (4b), respectively: $0 < b < 1$ implies $1 < \frac{1}{b}$ by (3.9). Now apply (4c) to see that, as x grows larger and larger, $b^x = \left(\frac{1}{b}\right)^{-x}$ gets arbitrarily close to 0. Similarly, apply (4b) to see that, as x grows smaller and smaller (i.e., as x "grows larger and larger in the negative direction"), $b^x = \left(\frac{1}{b}\right)^{-x}$ grows larger without bound.

Furthermore, (6) follows by (4a) and (5a) since $x_0 < x_1$ implies $b^{x_0} < b^{x_1}$ or $b^{x_0} > b^{x_1}$ and thus $b^{x_0} \neq b^{x_1}$.

Finally, (7) can be seen from (4b) and (4c) for $b > 1$, and from (5b) and (5c) for $0 < b < 1$, respectively.[8]

There is one issue we need to address still: How does the value of the base b impact on the graph of the exponential function? Figure 6.3 shows some examples; note that for clarity, we used various dotted lines for some of the graphs so you can quickly tell them apart. As you can see from these graphs, for $b > 1$, the bigger b is, the more steeply the graph of the exponential function increases. On the other hand, for $0 < b < 1$, the smaller b, the more steeply the graph of the exponential function decreases.

One final note: If you look at other algebra books, or at calculus books, you will find a special exponential function being mentioned, often called "the" exponential function. This is the exponential function $f(x) = e^x$ with special base $e \approx 2.718$ (which is sometimes called *Euler's number*). The fact that this base is special does not become apparent until one studies calculus; and using e as a base before is really more of a pain than a convenience since e is actually an irrational number. Therefore, we have chosen to ignore the base e here.

6.3.2. Another Application of Exponential Functions. We have already seen one application of exponential functions in "real life", namely, to bacterial growth in section 6.2. We will briefly cover one other application here to show you that exponential functions are indeed very useful, and not only in mathematics.

We had seen in section 6.2 that the "multiplicative change" of the number of bacteria from one hour to the next was constant: After each hour, the number of bacteria doubled. There is a much more "everyday" phenomenon in which such growth occurs, namely, in interest computations.

[8]Technically, we are cheating a bit here: The problem is that we haven't shown that there are no "holes" in the range of the exponential function; i.e., that the range really is the whole interval $(0, \infty)$. But calculus is required to actually show this carefully; so we're skipping over this fine point here.

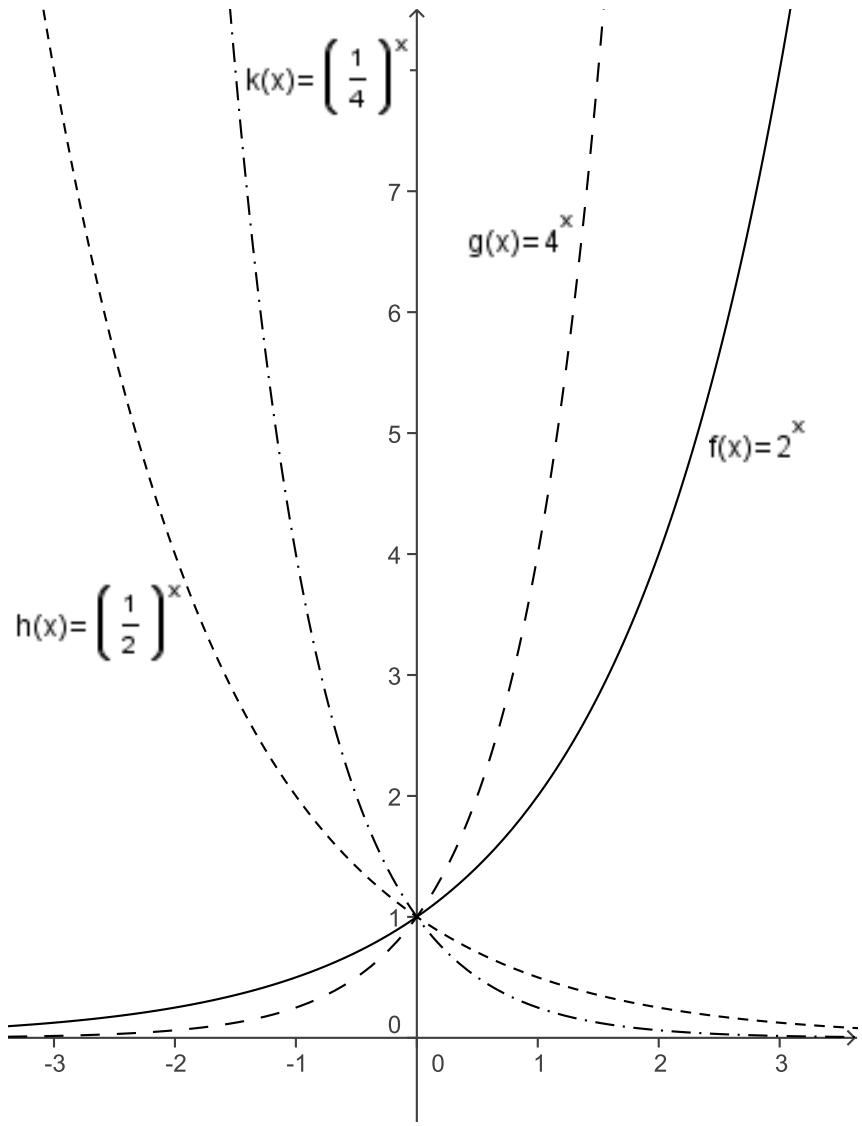

FIGURE 6.3. The graphs of various exponential functions

The setup for interest computations is typically as follows: You have a certain amount of capital (or debt) called the *principal*. At the end of each fixed time period (say, a year, or a month, or a quarter), a fixed percentage of the current amount is added. This added amount is called the *interest*, and the percentage is called the *interest rate* (per time period, e.g., per year). Typically, the interest is *compounded*, i.e.,

the interest is added to the amount, and next time interest is paid, interest is also added for the previous interest.

For a simple example, suppose I have \$100 in a savings account, and the bank pays me 5% interest at the end of each year. At the end of the first year, I then have \$105 (namely, the principal of \$100 plus 5% of it in interest); at the end of the second year, I have \$110.25 (namely, the previous amount of \$105 plus 5% of it in interest); etc. See Table 6.3.

number of years	0	1	2	3	4
amount in account	\$100.00	\$105.00	\$110.25	\$115.76	\$121.55

TABLE 6.3. Savings after n years

We can now see that my savings grow by a factor of $105\% = 1.05$ each year. Thus the capital I have is described by the function

$$C(n) = 100 \cdot 1.05^n \tag{6.17}$$

where n is the number of years I have saved. (This number n must be a whole number here for the problem to make sense.) More generally, the capital (or debt), starting with a principal P and under an interest rate of $r\%$ per time period, after n time periods, is given by the function

$$C(n) = P \cdot \left(1 + \frac{r}{100}\right)^n \tag{6.18}$$

where again the number n of time periods must be a whole number. Of course, (6.18) can also be used to solve for the principal P:

$$P = \frac{C}{\left(1 + \frac{r}{100}\right)^n}$$

or to solve for the interest rate r of time periods:

$$\begin{aligned} \left(1 + \frac{r}{100}\right)^n &= P/C \\ 1 + \frac{r}{100} &= \sqrt[n]{P/C} \\ r &= \sqrt[n]{P/C} - 1 \end{aligned}$$

It is also possible to solve for the number n of time periods, but that requires the logarithmic functions to be introduced in section 6.4.

Another application of exponential functions concerns nuclear decay, but again this will first require logarithmic functions and will be covered in subsection 6.4.2.

6.4. The Logarithmic Functions and Solving Exponential Equations

Our goal in this section is to find functions which allow us to solve *exponential equations*, i.e., equations of the form

$$b^x = c \tag{6.19}$$

where b and c are parameters. Compare this to previous situations where we tried to solve equations of the form

$$b + x = c \text{ or} \tag{6.20}$$

$$b \cdot x = c \tag{6.21}$$

We could easily solve (6.20) and (6.21) by "undoing" the operations of addition and multiplication, respectively:

$$x = c - b$$

$$x = \frac{c}{b}$$

In each case, we "undid" the equation by using an "inverse" function, namely, subtraction to "invert" addition, and division to "undo" multiplication.

More formally: In (6.20) and (6.21), we consider b to be fixed, so the functions we are inverting are the two functions $f(x) = b + x$ and $g(x) = b \cdot x$; and the inverse functions are $f^{-1}(y) = y - b$ and $g^{-1}(y) = \frac{y}{b}$, respectively. (Of course, when inverting g, we have to assume that $b \neq 0$; otherwise the function $g(x) = b \cdot x$ cannot be inverted.)

This should lead us to the conclusion that in order to solve exponential equations as in (6.19), we need to find an inverse function for the exponential function $f(x) = b^x$. We saw already that we need to assume that $b > 0$ for the exponential function to be defined. Also, we need to assume that $b \neq 1$; otherwise, the exponential function $f(x) = b^x$ is a constant function and cannot be inverted. So, for the rest of this section, we'll assume that $b > 0$ and $b \neq 1$.

6.4.1. Definition and Properties of Logarithmic Functions. In Proposition 6.3, we observed that for any positive base $b \neq 1$, the exponential function is a 1–1 function with domain $\mathbf{R}$ and range $(0, \infty)$. Typically, one sets the codomain (as defined in section 4.1) to be the set $(0, \infty)$ of positive reals so that the exponential function then becomes a 1–1 and onto function:

$$\begin{aligned} \exp_b : \mathbf{R} &\to (0, \infty) \\ x &\mapsto b^x \end{aligned} \tag{6.22}$$

Here, we have introduced the special symbol $\exp_b$ for the exponential function with base b.

Now recall from subsection 4.2.3 that a function has an inverse function exactly when the function is 1–1 and onto! This inverse function of the exponential function $\exp_b(x) = b^x$ is usually called the *logarithmic function to base b*; it is defined as

$$\begin{aligned} \log_b : (0, \infty) &\to \mathbf{R} \\ y &\mapsto \text{the unique } x \text{ such that } b^x = y \end{aligned} \tag{6.23}$$

Let's immediately return to solving exponential equations for a minute: The equation

$$b^x = c \tag{6.19}$$

can be solved using the logarithmic function $\log_b$ to arrive at

$$x = \log_b(c) \tag{6.24}$$

But how do we evaluate $\log_b(c)$ in practice, with actual numbers? E.g., for the bacterial growth example from section 6.2, in order to solve

$$2^n = 64 \tag{6.25}$$

we have to evaluate

$$n = \log_2(64) \tag{6.26}$$

Your first reaction may be: How do I evaluate $\log_2(64)$ on my calculator? There is no $\log_2$ button on my calculator? The only buttons I see are called "log" and "ln". (And that is only if you have a scientific calculator, and even then it may not have both.)

I'll give two answers here. First of all, in simple cases, you can compute logarithmic functions "in your head". To solve (6.26), you could simply recall the definition of $\log_2(64)$:

$$\log_2(64) = \text{the unique } n \text{ such that "}2^n = 64\text{" holds} \tag{6.27}$$

This means that you have to find the right exponent n to make the statement $2^n = 64$ true. A moment's thought tells you that that exponent must be 6.

Of course, this method will not always work, so how do you do this with a calculator? The real problem is that there is a different logarithmic function $\log_b$ for each base b, and no calculator gives them all. (Your calculator *could* give you the option of choosing both b and x when computing $\log_b(x)$, but I've never seen any calculator like that!)

Fortunately, there is a simple little trick. Most scientific calculators provide you with two special logarithmic functions, namely, log and ln.

The first, called log, is really $\log_{10}$ and sometimes called the *common logarithm* or the *decadic logarithm*. Its importance is mainly historic, as we will explain later on; but since it's somewhat easier than the other logarithm on your calculator, we will mainly use it. The other logarithmic function on your calculator, called ln, is really $\log_e$ for the number $e \approx 2.718$ mentioned at the end of section 6.3.1 and is usually called the *natural logarithm* (which explains the symbol ln); it is somewhat cumbersome and not all that useful until calculus (when it becomes *very* important, so important that most mathematicians call it "the" logarithmic function and ignore all others!).[9]

There is now a handy little formula which allows you to compute any logarithmic function from $\log_{10}$:

$$\log_b(x) = \frac{\log_{10}(x)}{\log_{10}(b)} \tag{6.28}$$

This formula is somewhat tedious to deduce, so we'll skip this here.[10] (Incidentally, the same formula also works with e in place of 10, i.e., we also have $\log_b(x) = \frac{\ln(x)}{\ln(b)}$.)

Note that formula (6.28) is exactly what you're looking for with a calculator which can handle $\log_{10}$: The right-hand side only involves computing $\log_{10}$ (and division). E.g., here is how we can compute $\log_2(64)$ the hard way, using a calculator:

$$\log_2(64) = \frac{\log_{10}(64)}{\log_{10}(2)} \approx \frac{1.806}{0.301} = 6$$

We will see applications of the logarithmic functions in subsection 6.4.2 where the use of a calculator (and thus the use of formula (6.28)) is inevitable.

But before we head into applications of the logarithmic functions, let's look at some properties of these functions. We summarize them in the following proposition, which is the counterpart of Proposition 6.3 for the exponential functions. (Figure 6.4 shows the graphs of some logarithmic functions and of the corresponding exponential functions;

[9]Some calculators have only one button for logarithmic functions, typically called log, and now you don't necessarily know which of the two logarithmic functions above that means, so be careful! You can test them as follows: If $\log(10)$ computes as 1 on your calculator, then you have $\log_{10}$, since $10^1 = 10$. If it computes as approximately 2.3, then you have $\log_e$, since $e^{2.3} \approx 2.7^{2.3} \approx 10$.

[10]But if you insist, here is the deduction: First of all, $x = b^{\log_b(x)} = \left(10^{\log_{10}(b)}\right)^{\log_b(x)}$ as well as $x = 10^{\log_{10}(x)}$ by definition of $\log_b$ and $\log_{10}$, respectively. By (6.12), this gives $10^{\log_{10}(b)\cdot\log_b(x)} = \left(10^{\log_{10}(b)}\right)^{\log_b(x)} = 10^{\log_{10}(x)}$. Since $\exp_{10}$ is 1–1, we obtain $\log_{10}(b)\cdot\log_b(x) = \log_{10}(x)$ and thus $\log_b(x) = \frac{\log_{10}(x)}{\log_{10}(b)}$.

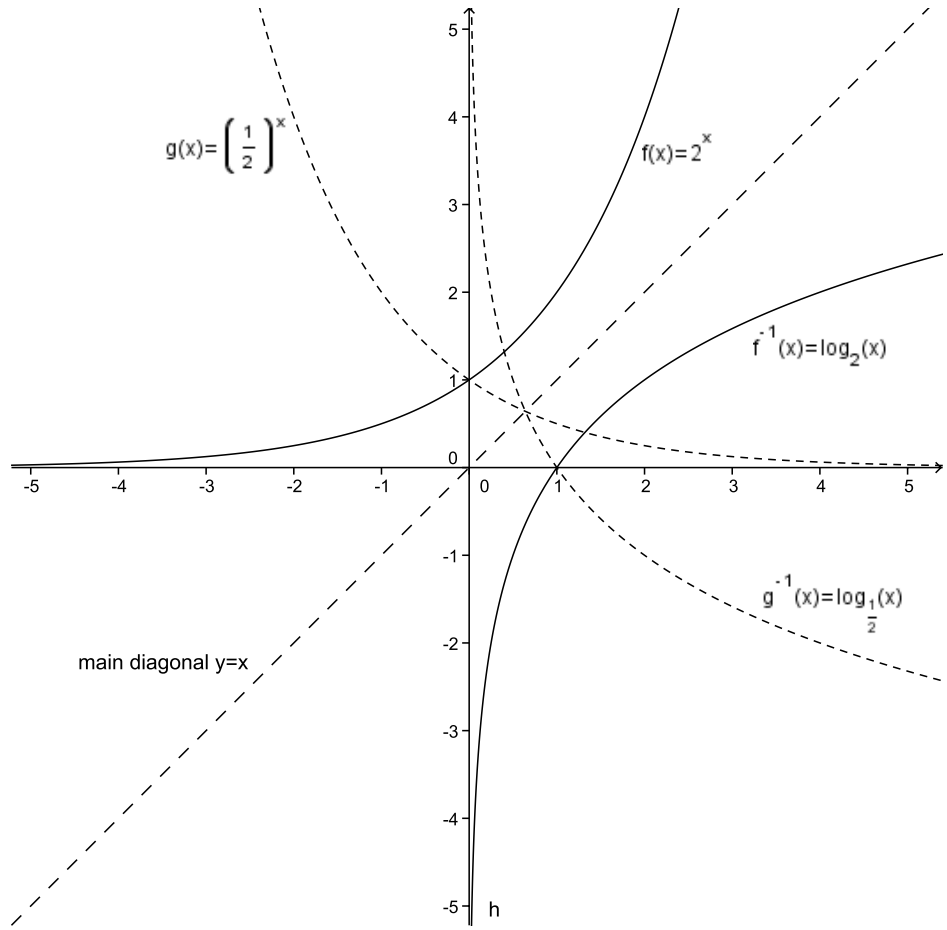

FIGURE 6.4. The graphs of two exponential functions and the corresponding logarithmic functions

note that for fixed base b, the graphs of the exponential and the logarithmic functions to base b are the reflections of each other along the "main diagonal" $y = x$ as mentioned in subsection 4.2.3. In order to better distinguish the graphs of the various functions in Figure 6.4, we have drawn some of them with "dotted" lines; but note that in reality, they should all be drawn with "solid" lines.)

PROPOSITION 6.4. *For any base $b > 0$ (with $b \neq 1$), the logarithmic function $\log_b$ has the following properties:*

(1) The logarithmic function $\log_b(x)$ is defined for all positive real numbers x, i.e., the domain of $\log_b$ is the set $(0, \infty)$ of all positive real numbers.
(2) The graph crosses the x-axis at the point $(0, 1)$.

(3) If $b > 1$ then
  (a) the logarithmic function $\log_b(x)$ increases as x increases, i.e., if $x_0 < x_1$ then $\log_b(x_0) < \log_b(x_1)$;
  (b) the graph increases "without bound" as x grows larger and larger; and
  (c) the graph decreases "without bound" as x approaches 0.

(4) If $0 < b < 1$ then
  (a) the logarithmic function $log_b(x)$ decreases as x increases, i.e., if $x_0 < x_1$ then $\log_b(x_0) > \log_b(x_1)$;
  (b) the graph decreases "without bound" as x grows larger and larger; and
  (c) the graph increases "without bound" as x approaches 0.

(5) The logarithmic function $\log_b$ is 1–1.

(6) The range of the logarithmic function $\log_b$ is the set $\mathbf{R}$ of all real numbers.

In summary, the logarithmic function $\log_b$ is a 1–1 function with domain $(0, \infty)$ (the set of all positive real numbers) and range $\mathbf{R}$.

The various parts of Proposition 6.4 can all be verified similarly to the way we verified Proposition 6.3; we'll skip this here for brevity's sake.

We finally note the following useful properties of the logarithmic functions, which follow from properties of the corresponding exponential functions stated in Proposition 6.2:

PROPOSITION 6.5. *The following* logarithmic rules *hold for all positive real numbers $a \neq 1$, all positive real numbers r and s, and all real numbers y:*

$$\log_a(r \cdot s) = \log_a(r) + \log_a(s) \tag{6.29}$$

$$\log_a(\frac{r}{s}) = \log_a r) - \log_a(s \tag{6.30}$$

$$\log_a(r^y) = y \cdot \log_a(r) \tag{6.31}$$

We will not prove these here, but merely list them for reference.[11]

[11]They actually are not hard to show: In each case, we set $r = a^x$ and $s = a^y$. Firstly, $\log_a(r \cdot s) = \log_a(r) + \log_a(s)$ holds since $a^x \cdot a^y = a^{x+y}$. Next, $\log_a(\frac{r}{s}) = \log_a(r) - \log_a(s)$ holds since $\frac{a^x}{a^y} = a^{x-y}$. And finally, $\log_a(r^y) = y \cdot \log_a(r)$ holds since $(a^x)^y = a^{xy}$.

6.4.2. Applications of Logarithmic Functions. Let's now finish our coverage of the logarithmic functions by looking at some applications. The problem with these is not that they are particularly difficult, but that each is typically phrased quite differently, which tends to confuse folks quite a bit.

6.4.2.1. *The slide rule.* The *slide rule* was invented in the 1600's as a mechanical calculator and was in wide use until the 1970's, when it was quickly crowded out by the modern pocket calculator.

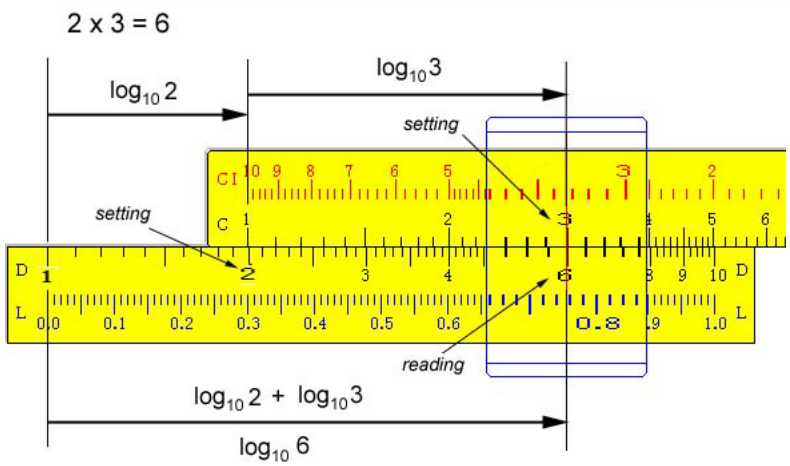

FIGURE 6.5. A slide rule

The basic idea of the slide rule is to take advantage of the logarithmic rules (6.29) and (6.30) in Proposition 6.5 in order to make (approximate) manual computations easier by converting multiplication to "graphical addition", and division to "graphical subtraction".[12]

Figure 6.5 shows a basic slide rule with a magnifying sliding cursor, allowing you to better align numbers on the scales; we will mainly use the middle two scales of it (called scales C and D on our slide rule) for now. Note that these two scales are sliding number lines, each from 1 to 10, but *not* drawn to scale like a usual number line, but such that each number x between 1 and 10 is drawn at a distance $\log_{10}(x)$ to the right of 1. Note, therefore, that the distance from 1 to 4 is twice the distance from 1 to 2 (since $\log_{10}(4) = \log_{10}(2^2) = 2 \cdot \log_{10}(2)$ by (6.31)), and similarly the distance from 1 to 8 is three times the distance from 1

[12]This short subsection is *not* intended as a full-blown guide to using a slide rule, but rather to give you the basic idea as to how and why a slide rule works. You can practice with a virtual slide rule on your computer screen at the web site `http://www.antiquark.com/sliderule/sim/n909es/virtual-n909-es.html`.

to 2 (since $\log_{10}(8) = \log_{10}(2^3) = 3 \cdot \log_{10}(2)$ by (6.31)). (You can check this against scale L at the very bottom, which *is* the usual number line from 0.0 to 1.0; its numbers represent the (base-10) logarithms of the numbers on scale D, so $\log_{10}(1) = 0$, $\log_{10}(2) \approx 0.3$, $\log_{10}(4) \approx 0.6$, $\log_{10}(8) \approx 0.9$, and $\log_{10}(10) = 1$.)

To illustrate the use of the slide rule, let's start with a simple multiplication problem: Suppose you want to multiply 2 by 3. The slide rule would accomplish this in the following steps (see Figure 6.5):

(1) Convert 2 and 3 to $\log_{10}(2)$ and $\log_{10}(3)$, represented by the length 2 on scale D and the length 3 on scale C, respectively.
(2) Add these two lengths and read off their sum as the length $\log_{10}(2) + \log_{10}(3) = \log_{10}(2 \cdot 3) = \log_{10}(6)$ on scale D (using (6.29)).
(3) Convert $\log_{10}(6)$ back to $10^{\log_{10}(6)} = 6$.

Similarly, in order to divide 6 by 3, say, we would proceed as follows (again refer to Figure 6.5):

(1) Convert 6 and 3 to $\log_{10}(6)$ and $\log_{10}(3)$, represented by the length 6 on scale D and the length 3 on scale C, respectively.
(2) Subtract these two lengths and read off their difference as the length $\log_{10}(6) - \log_{10}(3) = \log_{10}(\frac{6}{3}) = \log_{10}(2)$ on scale D (using (6.30)).
(3) Convert $\log_{10}(2)$ back to $10^{\log_{10}(2)} = 2$.

There are many other computations which can be performed on a slide rule, but since pocket calculators are nowadays a much more efficient way to perform the same computations, we'll leave it at that.

6.4.2.2. *Radioactive Decay.* The next application concerns radioactive materials: As radioactive material decays, its mass shrinks by a fixed amount per time period which is proportional to the amount left. Mathematically, this is similar to the balance of an interest-bearing savings account, which grows by a fixed amount per time period, as we discussed in subsection 6.3.2, except that the amount of radioactive material *decreases*, whereas the balance of an interest-bearing savings account *increases*.

Now, just to confuse you, the decay of a radioactive material is typically described by giving its *half-life*, namely, the time period it takes for the amount of material to decrease to half of its original amount. E.g., the half-life of plutonium is approximately 24,000 years; Table 6.4 shows the amount left after multiples of 24,000 years, starting with 1kg of plutonium. As you can see, the mass of the plutonium shrinks in half every 24,000 years, as the name "half-life" indicates.

time (years)	0	24,000	48,000	72,000	96,000	t
mass (kg)	1	0.5	0.25	0.125	0.0625	$\left(\frac{1}{2}\right)^{\frac{t}{24,000}}$

TABLE 6.4. Radioactive Decay: Approximate mass of plutonium left, starting with 1kg

But suppose we want to describe the amount of plutonium left after only 1,000 years? Well, after t many years, $\frac{t}{24,000}$ many multiples of the half-life period 24,000 years will have passed, so there will be $\left(\frac{1}{2}\right)^{\frac{t}{24,000}}$ many kg of plutonium left. So, in particular, after 1,000 years, there will be $\left(\frac{1}{2}\right)^{\frac{1,000}{24,000}} = \left(\frac{1}{2}\right)^{\frac{1}{24}} \approx .9715\text{kg} = 971.5\text{g}$ of plutonium left.

Now the logarithmic function also allows us to solve the inverse problem: Suppose I want to know after how many years there will be 100g of plutonium left. This then amounts to solving the equation

$$\left(\frac{1}{2}\right)^{\frac{t}{24,000}} = 0.1 \tag{6.32}$$

which can be undone as

$$\frac{t}{24,000} = \log_{\frac{1}{2}}(0.1)$$

giving

$$\begin{aligned} t &= 24,000 \cdot \log_{\frac{1}{2}}(0.1) \\ &= 24,000 \cdot \frac{\log(0.1)}{\log(0.5)} \text{ (by (6.28))} \\ &\approx 24,000 \cdot \frac{-1}{-0.301} \\ &\approx 24,000 \cdot 3.32 \approx 79,700 \end{aligned}$$

So it will take approximately 79,700 years for 90% of the plutonium to decay (and thus only 10% of it to be left).

6.4.2.3. *Earthquakes and the Richter Scale.* Let's look another application of logarithms, which at first looks quite different, the *Richter scale* for earthquakes. This is done as follows: One first measures the

maximum amplitude A (i.e., how far the ground "swings" from its resting position) at a distance of 100km from the epicenter;[13] the amplitude is measured in μm (read: micrometer, which is equal to 0.001mm or 0.000001m).

The problem with amplitude is that one has to deal with very large numbers: The amplitude of a barely noticeable earthquake is about 100μm; the amplitude of a very destructive earthquake is as high as 100,000,000μm = 100m or more. As you can see, it's not very practical to use the amplitude; so one uses the (decadic) logarithm of the amplitude, called the *Richter magnitude* (or, more correctly, called *local magnitude*)

$$M_L = \log_{10}(A) \tag{6.33}$$

Thus a barely noticeable earthquake of amplitude 100μm has Richter magnitude $M_L = \log_{10}(100) = \log_{10}(10^2) = 2$. On the other hand, a very destructive earthquake of amplitude 100,000,000μm has Richter magnitude $M_L = \log_{10}(100,000,000) = \log_{10}(10^8) = 8$. So as long as the amplitude is just a power of 10, the Richter magnitude will be a whole number (namely, the number of 0's in A when A is expressed in μm), and the Richter scale was chosen to make the Richter magnitude a number roughly between 0 and 10. One more quirk: If you have an earthquake of Richter magnitude M_0, say, then an earthquake of Richter magnitude $M_0 + 1$ has ten times the amplitude: This is because the first earthquake has amplitude A_0, say, where $M_0 = \log_{10}(A_0)$ or $A_0 = 10^{M_0}$; so the earthquake of Richter magnitude $M_0 + 1$ has amplitude A_1, say, where $M_0 + 1 = \log_{10}(A_1)$ or $A_1 = 10^{M_0+1} = 10 \cdot 10^{M_0} = 10 \cdot A_0$. This explains why an earthquake of Richter magnitude $M_0 + 1$ is generally much more destructive than an earthquake of Richter magnitude M_0.

6.4.2.4. *Sound Volume and Decibels.* The (decadic) logarithm appears similarly in the measurement of *(sound) volume* or *sound intensity*. One first expresses volume in terms of *power* P (formally defined in physics as *energy per time unit*), where the volume, or power, is expressed in watts (abbreviated W). Again the problem is that commonly occurring sound volumes can range from 1W (barely audible) to 1,000,000,000,000 W (one trillion watts, where ear damage occurs even during short exposure). So it is convenient to define a more practical

[13]There are formulas which one can use to convert the amplitude measured at a different distance from the epicenter to the amplitude one would have measured at a distance of 100km; we will not worry about this here.

measure of volume (or sound intensity) as

$$L_{\mathrm{dB}} = 10 \cdot \log_{10}(P) \tag{6.34}$$

When expressed as L_{dB}, the unit of volume is *decibels*.[14]

Using the above examples, we then see that a barely audible sound of power 1W has volume $L_{\mathrm{dB}} = 10 \cdot \log_{10}(1) = 10 \cdot log_{10}(10^0) = 0$ decibels, whereas a damaging sound of power 1,000,000,000,000 W has volume $L_{\mathrm{dB}} = 10 \cdot \log_{10}(1,000,000,000,000) = 10 \cdot \log_{10}(10^{12}) = 10 \cdot 12 = 120$ decibels.

[14]There is also a unit called *bel*, which equals 10 decibels and is now rarely used. It was defined as $\log_{10}(P)$. In a sense, bels are more similar to the definition of the Richter magnitude, but for some reason, it is now common practice to use the unit decibels.

Bibliography

[1] Gross, Herbert I., *Algebra by Example: An Elementary Course*, D. C. Heath & Co., Lexington, Mass., 1978 (out of print).

[2] Milgram, R. James, *The Mathematics Pre-Service Teachers Need to Know*, available on-line at `ftp://math.stanford.edu/pub/papers/milgram/FIE-book.pdf`, 2005 (especially "Chapter 8: Discussion of Issues in the Algebra Course").

[3] Papick, Ira J., *Algebra Connections: Mathematics for Middle School Teachers*, Prentice Hall, Upper Saddle River, NJ, 2007.

[4] Parker, Thomas H. and Baldridge, Scott J.. *Elementary Mathematics for Teachers*, Sefton-Ash Publishing, Okemos, Mich., 2004.

[5] Szydlik, Jennifer and Koker, John, *Big Ideas in Mathematics for Future Middle Grades Teachers: Big Ideas in Algebra*, University of Wisconsin-Oshkosh (preprint).

[6] Wu, Hung-Hsi, *Introduction to School Algebra [DRAFT]*, available on-line at `http://math.berkeley.edu/~wu/Algebrasummary.pdf`, 2005.

Index

Abel, Niels Henrik, vii
absolute value, 86
additive change, 130
additive identity, law of the, 8
additive inverse, 5
additive inverse, law of the, 8
Al-Kitab al-Jabr wa-l-Muqabala, vi
algebraic expression, 12
  evaluating, 12
  identical, 13
  negative, 66
  nonzero, 45
  positive, 65
  rewriting, 12
algorithm, vi
argument (of a function), 74
associative law (of addition), 8
associative law (of multiplication), 8

bacterial growth, 128
base, 124
binomial laws, 98

Cardano, Gerolamo, vii
ceiling function, 83
change
  additive, 130
  multiplicative, 130
codomain (of a function), 74
coefficient, 102
common logarithm, 140
commutative law (of addition), 8
commutative law (of multiplication), 8
completing the square, 103
complex numbers, 16, 112
compounding, 136
constant, 11
converse (of a statement), 34
coordinate, 30
coordinate axis, 30
coordinate system, 30
cube (of a number), 124
cubic equation, vii

decadic logarithm, 140
decibels, 147
decimal
  non-repeating, 16
  repeating, 16
del Ferro, Scipione, vii
difference (of sets), 17
difference function
  first, 93
  second, 95
direct variation, 89
direction (of velocity), 23
discriminant, 105
distance (on the number line), 88
distributive law, 8
domain (of a function), 73, 74

e (Euler's number), 135
earthquake, 146
elimination method, 54
empty set, 16
equality, properties of, 61
equation, 13
  equivalent, 43
Euler's number, 135
exponent, 124
exponential equation, 131, 138
exponential function, 131
exponential rules, 128

exponentiation
  definition, 124

Ferrari, Lodovico, vii
floor function, 83
fractions, 7
function, 72, 74

gradient, *see also* slope
graph, 25, 31, 75
greatest integer function, 83

half-life, 144
higher algebra, vi
history of algebra, vii

identity, 13
imaginary unit i, 111
index (of exponentiation), 124
inequality, 59
integers, 4, 16
intercept, 34, 35
interest, 136
interest rate, 136
intersection, 17
interval, 16
inverse function, 79
inverse variation, 89
inversely proportional, 89

least integer function, 83
letters in algebra, 8
line equation
  general form, 37
  intercept form, 39
  point-slope form, 36
  slope-intercept form, 34, 35
  two-point form, 37
linear equation, 30
  general, 48
linear equations
  simultaneous, 52
linear expression, 29
linear function, 82
linear inequality, 64
logarithm
  common, 140
  decadic, 140
  natural, 140
logarithmic function, 139
logarithmic rules, 142

main diagonal, 80
map, 74
mapping, 74
measurement division, 8
Muhammad ibn Mūsā al-Khwārizmī, vi
multiplicative change, 130
multiplicative identity, law of the, 8
multiplicative inverse, 7
multiplicative inverse, law of the, 8

natural logarithm, 140
numerical expression, 8

one-to-one (1–1) function, 78
onto function, 78
origin, 30

pair
  ordered pair of reals, 30
parabola, 116
parameter, 10
partitive division, 8
pattern, 92
polynomial division, 106
principal, 136
proportional, 89

quadratic equation, vii, 98
  general solution, 105
quadratic formula, 105
quadratic function, 95
quadratic inequality, 121
quartic equation, vii
quintic equation, vii

range (of a function), 78
rate (of change), 24
  constant, 24
ratio, 21
rational numbers, 7, 16
real numbers, 16
redursion
  definition by, 94
Richter scale, 146
root, 126

school algebra, vi

set, 15
slide rule, 143
slope, 33, 35
solution set, 67
speed, 23
square (of a number), 124
square root, 109
"step" function, 83
substitution method, 54

Tartaglia, Niccolò Fontana, vii

union, 17
unknown, 9

value (of a function), 74
variable, 9
  dependent, 9, 72
  independent, 9, 72
velocity, 23
vertex, 116
Viète, François (Vieta), vii

whole numbers, 4, 16