

Chapter 1

What Is Number Theory?

Number theory is the study of the set of positive whole numbers

$$1, 2, 3, 4, 5, 6, 7, \dots,$$

which are often called the set of *natural numbers*. We will especially want to study the *relationships* between different sorts of numbers. Since ancient times, people have separated the natural numbers into a variety of different types. Here are some familiar and not-so-familiar examples:

odd	1, 3, 5, 7, 9, 11, ...
even	2, 4, 6, 8, 10, ...
square	1, 4, 9, 16, 25, 36, ...
cube	1, 8, 27, 64, 125, ...
prime	2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, ...
composite	4, 6, 8, 9, 10, 12, 14, 15, 16, ...
1 (modulo 4)	1, 5, 9, 13, 17, 21, 25, ...
3 (modulo 4)	3, 7, 11, 15, 19, 23, 27, ...
triangular	1, 3, 6, 10, 15, 21, ...
perfect	6, 28, 496, ...
Fibonacci	1, 1, 2, 3, 5, 8, 13, 21, ...

Many of these types of numbers are undoubtedly already known to you. Others, such as the “modulo 4” numbers, may not be familiar. A number is said to be congruent to 1 (modulo 4) if it leaves a remainder of 1 when divided by 4, and similarly for the 3 (modulo 4) numbers. A number is called triangular if that number of pebbles can be arranged in a triangle, with one pebble at the top, two pebbles in the next row, and so on. The Fibonacci numbers are created by starting with 1 and 1. Then, to get the next number in the list, just add the previous two. Finally, a number is perfect if the sum of all its divisors, other than itself, adds back up to the

original number. Thus, the numbers dividing 6 are 1, 2, and 3, and $1 + 2 + 3 = 6$. Similarly, the divisors of 28 are 1, 2, 4, 7, and 14, and

$$1 + 2 + 4 + 7 + 14 = 28.$$

We will encounter all these types of numbers, and many others, in our excursion through the Theory of Numbers.

Some Typical Number Theoretic Questions

The main goal of number theory is to discover interesting and unexpected relationships between different sorts of numbers and to prove that these relationships are true. In this section we will describe a few typical number theoretic problems, some of which we will eventually solve, some of which have known solutions too difficult for us to include, and some of which remain unsolved to this day.

Sums of Squares I. Can the sum of two squares be a square? The answer is clearly “YES”; for example $3^2 + 4^2 = 5^2$ and $5^2 + 12^2 = 13^2$. These are examples of *Pythagorean triples*. We will describe all Pythagorean triples in Chapter 2.

Sums of Higher Powers. Can the sum of two cubes be a cube? Can the sum of two fourth powers be a fourth power? In general, can the sum of two n^{th} powers be an n^{th} power? The answer is “NO.” This famous problem, called *Fermat’s Last Theorem*, was first posed by Pierre de Fermat in the seventeenth century, but was not completely solved until 1994 by Andrew Wiles. Wiles’s proof uses sophisticated mathematical techniques that we will not be able to describe in detail, but in Chapter 30 we will prove that no fourth power is a sum of two fourth powers, and in Chapter 46 we will sketch some of the ideas that go into Wiles’s proof.

Infinitude of Primes. A *prime number* is a number p whose only factors are 1 and p .

- Are there infinitely many prime numbers?
- Are there infinitely many primes that are 1 modulo 4 numbers?
- Are there infinitely many primes that are 3 modulo 4 numbers?

The answer to all these questions is “YES.” We will prove these facts in Chapters 12 and 21 and also discuss a much more general result proved by Lejeune Dirichlet in 1837.

Sums of Squares II. Which numbers are sums of two squares? It often turns out that questions of this sort are easier to answer first for primes, so we ask which (odd) prime numbers are a sum of two squares. For example,

$$\begin{array}{llll} 3 = \text{NO}, & 5 = 1^2 + 2^2, & 7 = \text{NO}, & 11 = \text{NO}, \\ 13 = 2^2 + 3^2, & 17 = 1^2 + 4^2, & 19 = \text{NO}, & 23 = \text{NO}, \\ 29 = 2^2 + 5^2, & 31 = \text{NO}, & 37 = 1^2 + 6^2, & \dots \end{array}$$

Do you see a pattern? Possibly not, since this is only a short list, but a longer list leads to the conjecture that p is a sum of two squares if it is congruent to 1 (modulo 4). In other words, p is a sum of two squares if it leaves a remainder of 1 when divided by 4, and it is not a sum of two squares if it leaves a remainder of 3. We will prove that this is true in Chapter 24.

Number Shapes. The square numbers are the numbers 1, 4, 9, 16, ... that can be arranged in the shape of a square. The triangular numbers are the numbers 1, 3, 6, 10, ... that can be arranged in the shape of a triangle. The first few triangular and square numbers are illustrated in Figure 1.1.

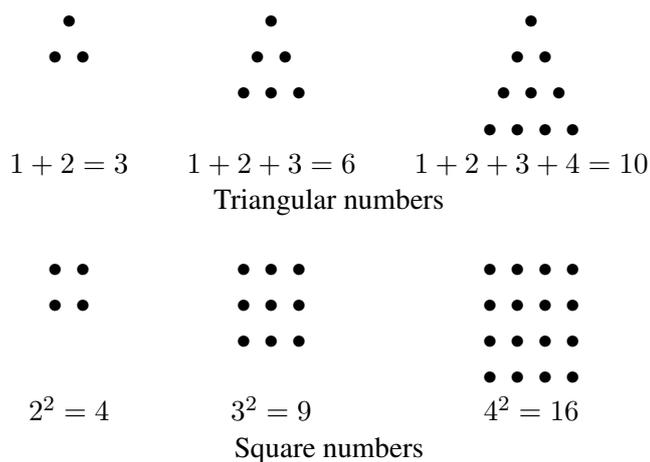


Figure 1.1: Numbers That Form Interesting Shapes

A natural question to ask is whether there are any triangular numbers that are also square numbers (other than 1). The answer is “YES,” the smallest example being

$$36 = 6^2 = 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8.$$

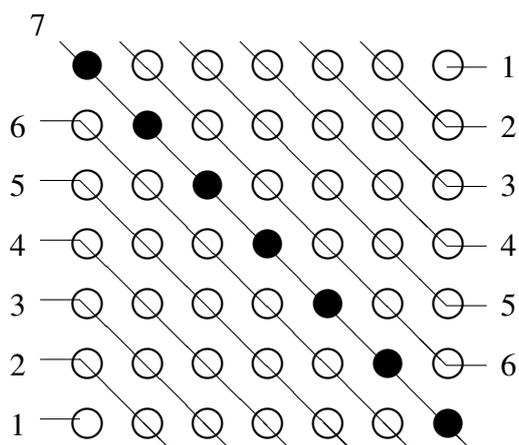
So we might ask whether there are more examples and, if so, are there in-

finitely many? To search for examples, the following formula is helpful:

$$1 + 2 + 3 + \cdots + (n - 1) + n = \frac{n(n + 1)}{2}.$$

There is an amusing anecdote associated with this formula. One day when the young Carl Friedrich Gauss (1777–1855) was in grade school, his teacher became so incensed with the class that he set them the task of adding up all the numbers from 1 to 100. As Gauss's classmates dutifully began to add, Gauss walked up to the teacher and presented the answer, 5050. The story goes that the teacher was neither impressed nor amused, but there's no record of what the next make-work assignment was!

There is an easy geometric way to verify Gauss's formula, which may be the way he discovered it himself. The idea is to take two triangles consisting of $1 + 2 + \cdots + n$ pebbles and fit them together with one additional diagonal of $n + 1$ pebbles. Figure 1.2 illustrates this idea for $n = 6$.



$$(1 + 2 + 3 + 4 + 5 + 6) + 7 + (6 + 5 + 4 + 3 + 2 + 1) = 7^2$$

Figure 1.2: The Sum of the First n Integers

In the figure, we have marked the extra $n + 1 = 7$ pebbles on the diagonal with black dots. The resulting square has sides consisting of $n + 1$ pebbles, so in mathematical terms we obtain the formula

$$2(1 + 2 + 3 + \cdots + n) + (n + 1) = (n + 1)^2,$$

two triangles + diagonal = square.

Now we can subtract $n + 1$ from each side and divide by 2 to get Gauss's formula.

Twin Primes. In the list of primes it is sometimes true that consecutive odd numbers are both prime. We have boxed these *twin primes* in the following list of primes less than 100:

$$\boxed{3}, \boxed{5}, \boxed{7}, \quad \boxed{11}, \boxed{13}, \quad \boxed{17}, \boxed{19}, \quad 23, \quad \boxed{29}, \boxed{31}, \quad 37$$

$$\boxed{41}, \boxed{43}, \quad 47, 53, \quad \boxed{59}, \boxed{61}, \quad 67, \quad \boxed{71}, \boxed{73}, \quad 79, 83, 89, 97.$$

Are there infinitely many twin primes? That is, are there infinitely many prime numbers p such that $p + 2$ is also a prime? At present, no one knows the answer to this question.

Primes of the Form $N^2 + 1$. If we list the numbers of the form $N^2 + 1$ taking $N = 1, 2, 3, \dots$, we find that some of them are prime. Of course, if N is odd, then $N^2 + 1$ is even, so it won't be prime unless $N = 1$. So it's really only interesting to take even values of N . We've highlighted the primes in the following list:

$$2^2 + 1 = \mathbf{5} \quad 4^2 + 1 = \mathbf{17} \quad 6^2 + 1 = \mathbf{37} \quad 8^2 + 1 = 65 = 5 \cdot 13$$

$$10^2 + 1 = \mathbf{101} \quad 12^2 + 1 = 145 = 5 \cdot 29 \quad 14^2 + 1 = \mathbf{197}$$

$$16^2 + 1 = \mathbf{257} \quad 18^2 + 1 = 325 = 5^2 \cdot 13 \quad 20^2 + 1 = \mathbf{401}.$$

It looks like there are quite a few prime values, but if you take larger values of N you will find that they become much rarer. So we ask whether there are infinitely many primes of the form $N^2 + 1$. Again, no one presently knows the answer to this question.

We have now seen some of the types of questions that are studied in the Theory of Numbers. How does one attempt to answer these questions? The answer is that Number Theory is partly experimental and partly theoretical. The experimental part normally comes first; it leads to questions and suggests ways to answer them. The theoretical part follows; in this part one tries to devise an argument that gives a conclusive answer to the questions. In summary, here are the steps to follow:

1. Accumulate data, usually numerical, but sometimes more abstract in nature.
2. Examine the data and try to find patterns and relationships.
3. Formulate conjectures (i.e., guesses) that explain the patterns and relationships. These are frequently given by formulas.

4. Test your conjectures by collecting additional data and checking whether the new information fits your conjectures.
5. Devise an argument (i.e., a proof) that your conjectures are correct.

All five steps are important in number theory and in mathematics. More generally, the scientific method always involves at least the first four steps. Be wary of any purported “scientist” who claims to have “proved” something using only the first three. Given any collection of data, it’s generally not too difficult to devise numerous explanations. The true test of a scientific theory is its ability to predict the outcome of experiments that have not yet taken place. In other words, a scientific theory only becomes plausible when it has been tested against new data. This is true of all real science. In mathematics one requires the further step of a proof, that is, a logical sequence of assertions, starting from known facts and ending at the desired statement.

Exercises

1.1. The first two numbers that are both squares and triangles are 1 and 36. Find the next one and, if possible, the one after that. Can you figure out an efficient way to find triangular–square numbers? Do you think that there are infinitely many?

1.2. Try adding up the first few odd numbers and see if the numbers you get satisfy some sort of pattern. Once you find the pattern, express it as a formula. Give a geometric verification that your formula is correct.

1.3. The consecutive odd numbers 3, 5, and 7 are all primes. Are there infinitely many such “prime triplets”? That is, are there infinitely many prime numbers p such that $p + 2$ and $p + 4$ are also primes?

1.4. It is generally believed that infinitely many primes have the form $N^2 + 1$, although no one knows for sure.

(a) Do you think that there are infinitely many primes of the form $N^2 - 1$?

(b) Do you think that there are infinitely many primes of the form $N^2 - 2$?

(c) How about of the form $N^2 - 3$? How about $N^2 - 4$?

(d) Which values of a do you think give infinitely many primes of the form $N^2 - a$?

1.5. The following two lines indicate another way to derive the formula for the sum of the first n integers by rearranging the terms in the sum. Fill in the details.

$$\begin{aligned} 1 + 2 + 3 + \cdots + n &= (1 + n) + (2 + (n - 1)) + (3 + (n - 2)) + \cdots \\ &= (1 + n) + (1 + n) + (1 + n) + \cdots . \end{aligned}$$

How many copies of $n + 1$ are in there in the second line? You may need to consider the cases of odd n and even n separately. If that’s not clear, first try writing it out explicitly for $n = 6$ and $n = 7$.

1.6. For each of the following statements, fill in the blank with an easy-to-check criterion:

- (a) M is a triangular number if and only if _____ is an odd square.
- (b) N is an odd square if and only if _____ is a triangular number.
- (c) Prove that your criteria in (a) and (b) are correct.

Chapter 2

Pythagorean Triples

The Pythagorean Theorem, that “beloved” formula of all high school geometry students, says that the sum of the squares of the sides of a right triangle equals the square of the hypotenuse. In symbols,

$$a^2 + b^2 = c^2$$

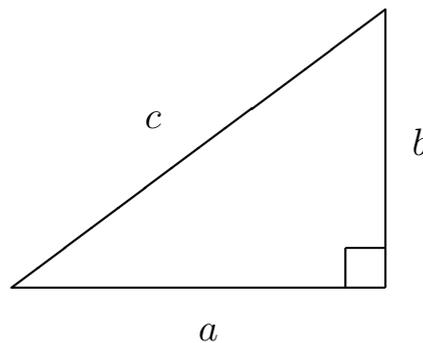


Figure 2.1: A Pythagorean Triangle

Since we’re interested in number theory, that is, the theory of the natural numbers, we will ask whether there are any Pythagorean triangles all of whose sides are natural numbers. There are many such triangles. The most famous has sides 3, 4, and 5. Here are the first few examples:

$$3^2 + 4^2 = 5^2, \quad 5^2 + 12^2 = 13^2, \quad 8^2 + 15^2 = 17^2, \quad 28^2 + 45^2 = 53^2.$$

The study of these *Pythagorean triples* began long before the time of Pythagoras. There are Babylonian tablets that contain lists of parts of such triples, including quite large ones, indicating that the Babylonians probably had a systematic method for producing them. Even more amazing is the fact that the Babylonians may have

used their lists of Pythagorean triples as primitive trigonometric tables. Pythagorean triples were also used in ancient Egypt. For example, a rough-and-ready way to produce a right angle is to take a piece of string, mark it into 12 equal segments, tie it into a loop, and hold it taut in the form of a 3-4-5 triangle, as illustrated in Figure 2.2. This provides an inexpensive right angle tool for use on small construction projects (such as marking property boundaries or building pyramids).

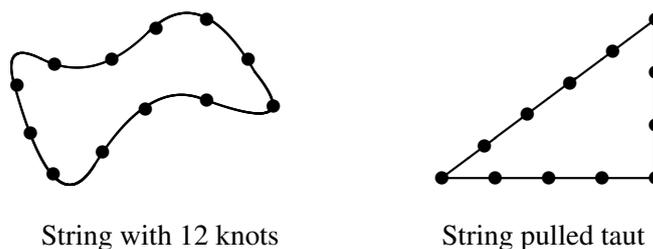


Figure 2.2: Using a knotted string to create a right triangle

The Babylonians and Egyptians had practical reasons for studying Pythagorean triples. Do such practical reasons still exist? For this particular problem, the answer is “probably not.” However, there is at least one good reason to study Pythagorean triples, and it’s the same reason why it is worthwhile studying the art of Rembrandt and the music of Beethoven. There is a beauty to the ways in which numbers interact with one another, just as there is a beauty in the composition of a painting or a symphony. To appreciate this beauty, one has to be willing to expend a certain amount of mental energy. But the end result is well worth the effort. Our goal in this book is to understand and appreciate some truly beautiful mathematics, to learn how this mathematics was discovered and proved, and maybe even to make some original contributions of our own.

Enough blathering, you are undoubtedly thinking. Let’s get to the real stuff. Our first naive question is whether there are infinitely many *Pythagorean triples*, that is, triples of natural numbers (a, b, c) satisfying the equation $a^2 + b^2 = c^2$. The answer is “YES” for a very silly reason. If we take a Pythagorean triple (a, b, c) and multiply it by some other number d , then we obtain a new Pythagorean triple (da, db, dc) . This is true because

$$(da)^2 + (db)^2 = d^2(a^2 + b^2) = d^2c^2 = (dc)^2.$$

Clearly these new Pythagorean triples are not very interesting. So we will concentrate our attention on triples with no common factors. We will even give them a name:

A *primitive Pythagorean triple* (or PPT for short) is a triple of numbers (a, b, c) such that a , b , and c have no common factors¹ and satisfy

$$a^2 + b^2 = c^2.$$

Recall our checklist from Chapter 1. The first step is to accumulate some data. I used a computer to substitute in values for a and b and checked if $a^2 + b^2$ is a square. Here are some primitive Pythagorean triples that I found:

$$\begin{array}{cccc} (3, 4, 5), & (5, 12, 13), & (8, 15, 17), & (7, 24, 25), \\ (20, 21, 29), & (9, 40, 41), & (12, 35, 37), & (11, 60, 61), \\ (28, 45, 53), & (33, 56, 65), & (16, 63, 65). \end{array}$$

A few conclusions can easily be drawn even from such a short list. For example, it certainly looks like one of a and b is odd and the other even. It also seems that c is always odd.

It's not hard to prove that these conjectures are correct. First, if a and b are both even, then c would also be even. This means that a , b , and c would have a common factor of 2, so the triple would not be primitive. Next, suppose that a and b are both odd, which means that c would have to be even. This means that there are numbers x , y , and z such that

$$a = 2x + 1, \quad b = 2y + 1, \quad \text{and} \quad c = 2z.$$

We can substitute these into the equation $a^2 + b^2 = c^2$ to get

$$\begin{aligned} (2x + 1)^2 + (2y + 1)^2 &= (2z)^2, \\ 4x^2 + 4x + 4y^2 + 4y + 2 &= 4z^2. \end{aligned}$$

Now divide by 2,

$$2x^2 + 2x + 2y^2 + 2y + 1 = 2z^2.$$

This last equation says that an odd number is equal to an even number, which is impossible, so a and b cannot both be odd. Since we've just checked that they cannot both be even and cannot both be odd, it must be true that one is even and

¹A *common factor* of a , b , and c is a number d such that each of a , b , and c is a multiple of d . For example, 3 is a common factor of 30, 42, and 105, since $30 = 3 \cdot 10$, $42 = 3 \cdot 14$, and $105 = 3 \cdot 35$, and indeed it is their largest common factor. On the other hand, the numbers 10, 12, and 15 have no common factor (other than 1). Since our goal in this chapter is to explore some interesting and beautiful number theory without getting bogged down in formalities, we will use common factors and divisibility informally and trust our intuition. In Chapter 5 we will return to these questions and develop the theory of divisibility more carefully.

the other is odd. It's then obvious from the equation $a^2 + b^2 = c^2$ that c is also odd.

We can always switch a and b , so our problem now is to find all solutions in natural numbers to the equation

$$a^2 + b^2 = c^2 \quad \text{with} \quad \begin{cases} a \text{ odd,} \\ b \text{ even,} \\ a, b, c \text{ having no common factors.} \end{cases}$$

The tools that we use are *factorization* and *divisibility*.

Our first observation is that if (a, b, c) is a primitive Pythagorean triple, then we can factor

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

Here are a few examples from the list given earlier, where note that we always take a to be odd and b to be even:

$$3^2 = 5^2 - 4^2 = (5 - 4)(5 + 4) = 1 \cdot 9,$$

$$15^2 = 17^2 - 8^2 = (17 - 8)(17 + 8) = 9 \cdot 25,$$

$$35^2 = 37^2 - 12^2 = (37 - 12)(37 + 12) = 25 \cdot 49,$$

$$33^2 = 65^2 - 56^2 = (65 - 56)(65 + 56) = 9 \cdot 121.$$

It looks like $c - b$ and $c + b$ are themselves always squares. We check this observation with a couple more examples:

$$21^2 = 29^2 - 20^2 = (29 - 20)(29 + 20) = 9 \cdot 49,$$

$$63^2 = 65^2 - 16^2 = (65 - 16)(65 + 16) = 49 \cdot 81.$$

How can we prove that $c - b$ and $c + b$ are squares? Another observation apparent from our list of examples is that $c - b$ and $c + b$ seem to have no common factors. We can prove this last assertion as follows. Suppose that d is a common factor of $c - b$ and $c + b$; that is, d divides both $c - b$ and $c + b$. Then d also divides

$$(c + b) + (c - b) = 2c \quad \text{and} \quad (c + b) - (c - b) = 2b.$$

Thus, d divides $2b$ and $2c$. But b and c have no common factor because we are assuming that (a, b, c) is a primitive Pythagorean triple. So d must equal 1 or 2. But d also divides $(c - b)(c + b) = a^2$, and a is odd, so d must be 1. In other words, the only number dividing both $c - b$ and $c + b$ is 1, so $c - b$ and $c + b$ have no common factor.

We now know that $c - b$ and $c + b$ are positive integers having no common factor, that their product is a square since $(c - b)(c + b) = a^2$. The only way that this can happen is if $c - b$ and $c + b$ are themselves squares.² So we can write

$$c + b = s^2 \quad \text{and} \quad c - b = t^2,$$

where $s > t \geq 1$ are odd integers with no common factors. Solving these two equations for b and c yields

$$c = \frac{s^2 + t^2}{2} \quad \text{and} \quad b = \frac{s^2 - t^2}{2},$$

and then

$$a = \sqrt{(c - b)(c + b)} = st.$$

We have (almost) finished our first proof! The following theorem records our accomplishment.

Theorem 2.1 (Pythagorean Triples Theorem). *We will get every primitive Pythagorean triple (a, b, c) with a odd and b even by using the formulas*

$$a = st, \quad b = \frac{s^2 - t^2}{2}, \quad c = \frac{s^2 + t^2}{2},$$

where $s > t \geq 1$ are chosen to be any odd integers with no common factors.

Why did we say that we have “almost” finished the proof? We have shown that if (a, b, c) is a PPT with a odd, then there are odd integers $s > t \geq 1$ with no common factors so that a , b , and c are given by the stated formulas. But we still need to check that these formulas always give a PPT. We first use a little bit of algebra to show that the formulas give a Pythagorean triple. Thus

$$(st)^2 + \left(\frac{s^2 - t^2}{2}\right)^2 = s^2t^2 + \frac{s^4 - 2s^2t^2 + t^4}{4} = \frac{s^4 + 2s^2t^2 + t^4}{4} = \left(\frac{s^2 + t^2}{2}\right)^2.$$

We also need to check that st , $\frac{s^2 - t^2}{2}$, and $\frac{s^2 + t^2}{2}$ have no common factors. This is most easily accomplished using an important property of prime numbers, so we postpone the proof until Chapter 7, where you will finish the argument (Exercise 7.3).

²This is intuitively clear if you consider the factorization of $c - b$ and $c + b$ into primes, since the primes in the factorization of $c - b$ will be distinct from the primes in the factorization of $c + b$. However, the existence and uniqueness of the factorization into primes is by no means as obvious as it appears. We will discuss this further in Chapter 7.

For example, taking $t = 1$ in Theorem 2.1 gives a triple $\left(s, \frac{s^2-1}{2}, \frac{s^2+1}{2}\right)$ whose b and c entries differ by 1. This explains many of the examples that we listed. The following table gives all possible triples with $s \leq 9$.

s	t	$a = st$	$b = \frac{s^2 - t^2}{2}$	$c = \frac{s^2 + t^2}{2}$
3	1	3	4	5
5	1	5	12	13
7	1	7	24	25
9	1	9	40	41
5	3	15	8	17
7	3	21	20	29
7	5	35	12	37
9	5	45	28	53
9	7	63	16	65

A Notational Interlude

Mathematicians have created certain standard notations as a shorthand for various quantities. We will keep our use of such notation to a minimum, but there are a few symbols that are so commonly used and are so useful that it is worthwhile to introduce them here. They are

\mathbb{N} = the set of natural numbers = $1, 2, 3, 4, \dots$,

\mathbb{Z} = the set of integers = $\dots - 3, -2, -1, 0, 1, 2, 3, \dots$,

\mathbb{Q} = the set of rational numbers (i.e., fractions).

In addition, mathematicians often use \mathbb{R} to denote the real numbers and \mathbb{C} for the complex numbers, but we will not need these. Why were these letters chosen? The choice of \mathbb{N} , \mathbb{R} , and \mathbb{C} needs no explanation. The letter \mathbb{Z} for the set of integers comes from the German word “Zahlen,” which means numbers. Similarly, \mathbb{Q} comes from the German “Quotient” (which is the same as the English word). We will also use the standard mathematical symbol \in to mean “is an element of the set.” So, for example, $a \in \mathbb{N}$ means that a is a natural number, and $x \in \mathbb{Q}$ means that x is a rational number.

Exercises

- 2.1. (a)** We showed that in any primitive Pythagorean triple (a, b, c) , either a or b is even. Use the same sort of argument to show that either a or b must be a multiple of 3.

(b) By examining the above list of primitive Pythagorean triples, make a guess about when a , b , or c is a multiple of 5. Try to show that your guess is correct.

2.2. A nonzero integer d is said to *divide* an integer m if $m = dk$ for some number k . Show that if d divides both m and n , then d also divides $m - n$ and $m + n$.

2.3. For each of the following questions, begin by compiling some data; next examine the data and formulate a conjecture; and finally try to prove that your conjecture is correct. (But don't worry if you can't solve every part of this problem; some parts are quite difficult.)

- (a) Which odd numbers a can appear in a primitive Pythagorean triple (a, b, c) ?
- (b) Which even numbers b can appear in a primitive Pythagorean triple (a, b, c) ?
- (c) Which numbers c can appear in a primitive Pythagorean triple (a, b, c) ?

2.4. In our list of examples are the two primitive Pythagorean triples

$$33^2 + 56^2 = 65^2 \quad \text{and} \quad 16^2 + 63^2 = 65^2.$$

Find at least one more example of two primitive Pythagorean triples with the same value of c . Can you find three primitive Pythagorean triples with the same c ? Can you find more than three?

2.5. In Chapter 1 we saw that the n^{th} triangular number T_n is given by the formula

$$T_n = 1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}.$$

The first few triangular numbers are 1, 3, 6, and 10. In the list of the first few Pythagorean triples (a, b, c) , we find $(3, 4, 5)$, $(5, 12, 13)$, $(7, 24, 25)$, and $(9, 40, 41)$. Notice that in each case, the value of b is four times a triangular number.

- (a) Find a primitive Pythagorean triple (a, b, c) with $b = 4T_5$. Do the same for $b = 4T_6$ and for $b = 4T_7$.
- (b) Do you think that for every triangular number T_n , there is a primitive Pythagorean triple (a, b, c) with $b = 4T_n$? If you believe that this is true, then prove it. Otherwise, find some triangular number for which it is not true.

2.6. If you look at the table of primitive Pythagorean triples in this chapter, you will see many triples in which c is 2 greater than a . For example, the triples $(3, 4, 5)$, $(15, 8, 17)$, $(35, 12, 37)$, and $(63, 16, 65)$ all have this property.

- (a) Find two more primitive Pythagorean triples (a, b, c) having $c = a + 2$.
- (b) Find a primitive Pythagorean triple (a, b, c) having $c = a + 2$ and $c > 1000$.
- (c) Try to find a formula that describes all primitive Pythagorean triples (a, b, c) having $c = a + 2$.

2.7. For each primitive Pythagorean triple (a, b, c) in the table in this chapter, compute the quantity $2c - 2a$. Do these values seem to have some special form? Try to prove that your observation is true for all primitive Pythagorean triples.

2.8. Let m and n be numbers that differ by 2, and write the sum $\frac{1}{m} + \frac{1}{n}$ as a fraction in lowest terms. For example, $\frac{1}{2} + \frac{1}{4} = \frac{3}{4}$ and $\frac{1}{3} + \frac{1}{5} = \frac{8}{15}$.

-
- (a) Compute the next three examples.
 - (b) Examine the numerators and denominators of the fractions in (a) and compare them with the table of Pythagorean triples on page 18. Formulate a conjecture about such fractions.
 - (c) Prove that your conjecture is correct.
- 2.9.**
- (a) Read about the Babylonian number system and write a short description, including the symbols for the numbers 1 to 10 and the multiples of 10 from 20 to 50.
 - (b) Read about the Babylonian tablet called Plimpton 322 and write a brief report, including its approximate date of origin.
 - (c) The second and third columns of Plimpton 322 give pairs of integers (a, c) having the property that $c^2 - a^2$ is a perfect square. Convert some of these pairs from Babylonian numbers to decimal numbers and compute the value of b so that (a, b, c) is a Pythagorean triple.

Chapter 3

Pythagorean Triples and the Unit Circle

In the previous chapter we described all solutions to

$$a^2 + b^2 = c^2$$

in whole numbers a, b, c . If we divide this equation by c^2 , we obtain

$$\left(\frac{a}{c}\right)^2 + \left(\frac{b}{c}\right)^2 = 1.$$

So the pair of rational numbers $(a/c, b/c)$ is a solution to the equation

$$x^2 + y^2 = 1.$$

Everyone knows what the equation $x^2 + y^2 = 1$ looks like: It is a circle C of radius 1 with center at $(0, 0)$. We are going to use the geometry of the circle C to find all the points on C whose xy -coordinates are rational numbers. Notice that the circle has four obvious points with rational coordinates, $(\pm 1, 0)$ and $(0, \pm 1)$. Suppose that we take any (rational) number m and look at the line L going through the point $(-1, 0)$ and having slope m . (See Figure 3.1.) The line L is given by the equation

$$L : y = m(x + 1) \quad (\text{point-slope formula}).$$

It is clear from the picture that the intersection $C \cap L$ consists of exactly two points, and one of those points is $(-1, 0)$. We want to find the other one.

To find the intersection of C and L , we need to solve the equations

$$x^2 + y^2 = 1 \quad \text{and} \quad y = m(x + 1)$$

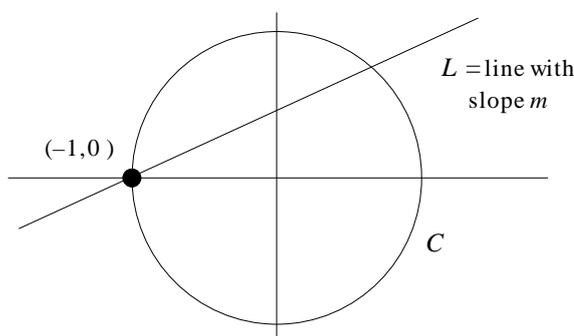


Figure 3.1: The Intersection of a Circle and a Line

for x and y . Substituting the second equation into the first and simplifying, we need to solve

$$\begin{aligned}x^2 + (m(x+1))^2 &= 1 \\x^2 + m^2(x^2 + 2x + 1) &= 1 \\(m^2 + 1)x^2 + 2m^2x + (m^2 - 1) &= 0.\end{aligned}$$

This is just a quadratic equation, so we could use the quadratic formula to solve for x . But there is a much easier way to find the solution. We know that $x = -1$ must be a solution, since the point $(-1, 0)$ is on both C and L . This means that we can divide the quadratic polynomial by $x + 1$ to find the other root:

$$x + 1 \overline{) \begin{array}{r} (m^2 + 1)x + (m^2 - 1) \\ (m^2 + 1)x^2 + 2m^2x + (m^2 - 1) \end{array}} .$$

So the other root is the solution of $(m^2 + 1)x + (m^2 - 1) = 0$, which means that

$$x = \frac{1 - m^2}{1 + m^2}.$$

Then we substitute this value of x into the equation $y = m(x + 1)$ of the line L to find the y -coordinate,

$$y = m(x + 1) = m \left(\frac{1 - m^2}{1 + m^2} + 1 \right) = \frac{2m}{1 + m^2}.$$

Thus, for every rational number m we get a solution in rational numbers

$$\left(\frac{1 - m^2}{1 + m^2}, \frac{2m}{1 + m^2} \right) \text{ to the equation } x^2 + y^2 = 1.$$

On the other hand, if we have a solution (x_1, y_1) in rational numbers, then the slope of the line through (x_1, y_1) and $(-1, 0)$ will be a rational number. So by taking all possible values for m , the process we have described will yield every solution to $x^2 + y^2 = 1$ in rational numbers [except for $(-1, 0)$, which corresponds to a vertical line having slope “ $m = \infty$ ”]. We summarize our results in the following theorem.

Theorem 3.1. *Every point on the circle*

$$x^2 + y^2 = 1$$

whose coordinates are rational numbers can be obtained from the formula

$$(x, y) = \left(\frac{1 - m^2}{1 + m^2}, \frac{2m}{1 + m^2} \right)$$

by substituting in rational numbers for m [except for the point $(-1, 0)$ which is the limiting value as $m \rightarrow \infty$].

How is this formula for rational points on a circle related to our formula for Pythagorean triples? If we write the rational number m as a fraction v/u , then our formula becomes

$$(x, y) = \left(\frac{u^2 - v^2}{u^2 + v^2}, \frac{2uv}{u^2 + v^2} \right),$$

and clearing denominators gives the Pythagorean triple

$$(a, b, c) = (u^2 - v^2, 2uv, u^2 + v^2).$$

This is another way of describing Pythagorean triples, although to describe only the primitive ones would require some restrictions on u and v . You can relate this description to the formula in Chapter 2 by setting

$$u = \frac{s+t}{2} \quad \text{and} \quad v = \frac{s-t}{2}.$$

Exercises

3.1. As we have just seen, we get every Pythagorean triple (a, b, c) with b even from the formula

$$(a, b, c) = (u^2 - v^2, 2uv, u^2 + v^2)$$

by substituting in different integers for u and v . For example, $(u, v) = (2, 1)$ gives the smallest triple $(3, 4, 5)$.

- (a) If u and v have a common factor, explain why (a, b, c) will not be a primitive Pythagorean triple.
- (b) Find an example of integers $u > v > 0$ that do not have a common factor, yet the Pythagorean triple $(u^2 - v^2, 2uv, u^2 + v^2)$ is not primitive.
- (c) Make a table of the Pythagorean triples that arise when you substitute in all values of u and v with $1 \leq v < u \leq 10$.
- (d) Using your table from (c), find some simple conditions on u and v that ensure that the Pythagorean triple $(u^2 - v^2, 2uv, u^2 + v^2)$ is primitive.
- (e) Prove that your conditions in (d) really work.

3.2. (a) Use the lines through the point $(1, 1)$ to describe all the points on the circle

$$x^2 + y^2 = 2$$

whose coordinates are rational numbers.

- (b) What goes wrong if you try to apply the same procedure to find all the points on the circle $x^2 + y^2 = 3$ with rational coordinates?

3.3. Find a formula for all the points on the hyperbola

$$x^2 - y^2 = 1$$

whose coordinates are rational numbers. [*Hint.* Take the line through the point $(-1, 0)$ having rational slope m and find a formula in terms of m for the second point where the line intersects the hyperbola.]

3.4. The curve

$$y^2 = x^3 + 8$$

contains the points $(1, -3)$ and $(-7/4, 13/8)$. The line through these two points intersects the curve in exactly one other point. Find this third point. Can you explain why the coordinates of this third point are rational numbers?

3.5. Numbers that are both square and triangular numbers were introduced in Chapter 1, and you studied them in Exercise 1.1.

- (a) Show that every square–triangular number can be described using the solutions in positive integers to the equation $x^2 - 2y^2 = 1$. [*Hint.* Rearrange the equation $m^2 = \frac{1}{2}(n^2 + n)$.]
- (b) The curve $x^2 - 2y^2 = 1$ includes the point $(1, 0)$. Let L be the line through $(1, 0)$ having slope m . Find the other point where L intersects the curve.
- (c) Suppose that you take m to equal $m = v/u$, where (u, v) is a solution to $u^2 - 2v^2 = 1$. Show that the other point that you found in (b) has integer coordinates. Further, changing the signs of the coordinates if necessary, show that you get a solution to $x^2 - 2y^2 = 1$ in positive integers.
- (d) Starting with the solution $(3, 2)$ to $x^2 - 2y^2 = 1$, apply (b) and (c) repeatedly to find several more solutions to $x^2 - 2y^2 = 1$. Then use those solutions to find additional examples of square–triangular numbers.

- (e) Prove that this procedure leads to infinitely many different square-triangular numbers.
- (f) Prove that every square-triangular number can be constructed in this way. (This part is very difficult. Don't worry if you can't solve it.)

Chapter 4

Sums of Higher Powers and Fermat's Last Theorem

In the previous two chapters we discovered that the equation

$$a^2 + b^2 = c^2$$

has lots of solutions in whole numbers a, b, c . It is natural to ask whether there are solutions when the exponent 2 is replaced by a higher power. For example, do the equations

$$a^3 + b^3 = c^3 \quad \text{and} \quad a^4 + b^4 = c^4 \quad \text{and} \quad a^5 + b^5 = c^5$$

have solutions in nonzero integers a, b, c ? The answer is “NO.” Sometime around 1637, Pierre de Fermat showed that there is no solution for exponent 4. During the eighteenth and nineteenth centuries, Carl Friedrich Gauss and Leonhard Euler showed that there is no solution for exponent 3 and Lejeune Dirichlet and Adrien Legendre dealt with the exponent 5. The general problem of showing that the equation

$$a^n + b^n = c^n$$

has no solutions in positive integers if $n \geq 3$ is known as “Fermat's Last Theorem.” It has attained almost cult status in the 350 years since Fermat scribbled the following assertion in the margin of one of his books:

It is impossible to separate a cube into two cubes, or a fourth power into two fourth powers, or in general any power higher than the second into powers of

like degree. I have discovered a truly remarkable proof which this margin is too small to contain.¹

Few mathematicians today believe that Fermat had a valid proof of his “Theorem,” which is called his Last Theorem because it was the last of his assertions that remained unproved. The history of Fermat's Last Theorem is fascinating, with literally hundreds of mathematicians making important contributions. Even a brief summary could easily fill a book. This is not our intent in this volume, so we will be content with a few brief remarks.

One of the first general results on Fermat's Last Theorem, as opposed to verification for specific exponents n , was given by Sophie Germain in 1823. She proved that if both p and $2p + 1$ are primes then the equation $a^p + b^p = c^p$ has no solutions in integers a, b, c with p not dividing the product abc . A later result of a similar nature, due to A. Wieferich in 1909, is that the same conclusion is true if the quantity $2^p - 2$ is not divisible by p^2 . Meanwhile, during the latter part of the nineteenth century a number of mathematicians, including Richard Dedekind, Leopold Kronecker, and especially Ernst Kummer, developed a new field of mathematics called algebraic number theory and used their theory to prove Fermat's Last Theorem for many exponents, although still only a finite list. Then, in 1985, L.M. Adleman, D.R. Heath-Brown, and E. Fouvry used a refinement of Germain's criterion together with difficult analytic estimates to prove that there are infinitely many primes p such that $a^p + b^p = c^p$ has no solutions with p not dividing abc .

Sophie Germain (1776–1831) Sophie Germain was a French mathematician who did important work in number theory and differential equations. She is best known for her work on Fermat's Last Theorem, where she gave a simple criterion that suffices to show that the equation $a^p + b^p = c^p$ has no solutions with abc not divisible by p . She also did work on acoustics and elasticity, especially the theory of vibrating plates. As a mathematics student, she was forced to take correspondence courses from the École Polytechnique in Paris, since they did not accept women as students. For a similar reason, she began her extensive correspondence with Gauss using the pseudonym Monsieur Le Blanc; but when she eventually revealed her identity, Gauss was delighted and sufficiently impressed with her work to recommend her for an honorary degree at the University of Göttingen.

In 1986 Gerhard Frey suggested a new line of attack on Fermat's problem using a notion called modularity. Frey's idea was refined by Jean-Pierre Serre, and Ken

¹Translated from the Latin: “*Cubum autem in duos cubos, aut quadrato quadratum in duos quadrato quadratos, & generaliter nullam in infinitum ultra quadratum potestatem in duos ejusdem nominis fas est dividere; cujus rei demonstrationem mirabilem sane detexi. Hanc marginis exiguitas non caperet.*”

Ribet subsequently proved that if the Modularity Conjecture is true, then Fermat's Last Theorem is true. Precisely, Ribet proved that if every semistable elliptic curve² is modular³ then Fermat's Last Theorem is true. The Modularity Conjecture, which asserts that every rational elliptic curve is modular, was at that time a conjecture originally formulated by Goro Shimura and Yutaka Taniyama. Finally, in 1994, Andrew Wiles announced a proof that every semistable rational elliptic curve is modular, thereby completing the proof of Fermat's 350-year-old claim. Wiles's proof, which is a tour de force using the vast machinery of modern number theory and algebraic geometry, is far too complicated for us to describe in detail, but we will try to convey the flavor of his proof in Chapter 46.

Few mathematical or scientific discoveries arise in a vacuum. Even Sir Isaac Newton, the transcendent genius not noted for his modesty, wrote that "If I have seen further, it is by standing on the shoulders of giants." Here is a list of some of the giants, all contemporary mathematicians, whose work either directly or indirectly contributed to Wiles's brilliant proof. The diversified nationalities highlight the international character of modern mathematics. In alphabetical order: Spencer Bloch (USA), Henri Carayol (France), John Coates (Australia), Pierre Deligne (Belgium), Ehud de Shalit (Israel), Fred Diamond (USA), Gerd Faltings (Germany), Matthias Flach (Germany), Gerhard Frey (Germany), Alexander Grothendieck (France), Yves Hellegouarch (France), Haruzo Hida (Japan), Kenkichi Iwasawa (Japan), Kazuya Kato (Japan), Nick Katz (USA), V.A. Kolyvagin (Russia), Ernst Kunz (Germany), Robert Langlands (Canada), Hendrik Lenstra (The Netherlands), Wen-Ch'ing Winnie Li (USA), Barry Mazur (USA), André Néron (France), Ravi Ramakrishna (USA), Michel Raynaud (France), Ken Ribet (USA), Karl Rubin (USA), Jean-Pierre Serre (France), Goro Shimura (Japan), Yutaka Taniyama (Japan), John Tate (USA), Richard Taylor (England), Jacques Tilouine (France), Jerry Tunnell (USA), André Weil (France), Andrew Wiles (England).

Exercises

4.1. Write a one- to two-page biography on one (or more) of the following mathematicians. Be sure to describe their mathematical achievements, especially in number theory, and some details of their lives. Also include a paragraph putting them into an historical context

²An elliptic curve is a certain sort of curve, not an ellipse, given by an equation of the form $y^2 = x^3 + ax^2 + bx + c$, where a, b, c are integers. The elliptic curve is semistable if the quantities $3b - a^2$ and $27c - 9ab + 2a^3$ have no common factors other than 2 and satisfy a few other technical conditions. We study elliptic curves in Chapters 41–46.

³An elliptic curve is called modular if there is a map to it from another special sort of curve called a modular curve.

by describing the times (scientifically, politically, socially, etc.) during which they lived and worked: (a) Niels Abel, (b) Claude Gaspar Bachet de Meziriac, (c) Richard Dedekind, (d) Diophantus of Alexandria, (e) Lejeune Dirichlet, (f) Eratosthenes, (g) Euclid of Alexandria, (h) Leonhard Euler, (i) Pierre de Fermat, (j) Leonardo Fibonacci, (k) Carl Friedrich Gauss, (l) Sophie Germain, (m) David Hilbert, (n) Carl Jacobi, (o) Leopold Kronecker, (p) Ernst Kummer, (q) Joseph-Louis Lagrange, (r) Adrien-Marie Legendre, (s) Joseph Liouville, (t) Marin Mersenne, (u) Hermann Minkowski, (v) Sir Isaac Newton, (w) Pythagoras, (x) Srinivasa Ramanujan, (y) Bernhard Riemann, (z) P.L. Tchebychef (also spelled Chebychev).

4.2. The equation $a^2 + b^2 = c^2$ has lots of solutions in positive integers, while the equation $a^3 + b^3 = c^3$ has no solutions in positive integers. This exercise asks you to look for solutions to the equation

$$a^3 + b^3 = c^2 \tag{*}$$

in integers $c \geq b \geq a \geq 1$.

- (a) The equation (*) has the solution $(a, b, c) = (2, 2, 4)$. Find three more solutions in positive integers. [*Hint.* Look for solutions of the form $(a, b, c) = (xz, yz, z^2)$. Not every choice of x, y, z will work, of course, so you'll need to figure out which ones do work.]
- (b) If (A, B, C) is a solution to (*) and n is any integer, show that (n^2A, n^2B, n^3C) is also a solution to (*). We will say that a solution (a, b, c) to (*) is *primitive* if it does not look like (n^2A, n^2B, n^3C) for any $n \geq 2$.
- (c) Write down four different primitive solutions to (*). [That is, redo (a) using only primitive solutions.]
- (d) The solution $(2, 2, 4)$ has $a = b$. Find all primitive solutions that have $a = b$.
- (e) Find a primitive solution to (*) that has $a > 10000$.

Chapter 5

Divisibility and the Greatest Common Divisor

As we have already seen in our study of Pythagorean triples, the notions of divisibility and factorizations are important tools in number theory. In this chapter we will look at these ideas more closely.

Suppose that m and n are integers with $m \neq 0$. We say that m divides n if n is a multiple of m , that is, if there is an integer k such that $n = mk$. If m divides n , we write $m|n$. Similarly, if m does not divide n , then we write $m \nmid n$. For example,

$$3|6 \quad \text{and} \quad 12|132, \quad \text{since} \quad 6 = 3 \cdot 2 \quad \text{and} \quad 132 = 12 \cdot 11.$$

The divisors of 6 are 1, 2, 3, and 6. On the other hand, $5 \nmid 7$, since no integer multiple of 5 is equal to 7. A number that divides n is called a *divisor of n* .

If we are given two numbers, we can look for common divisors, that is, numbers that divide both of them. For example, 4 is a common divisor of 12 and 20, since $4|12$ and $4|20$. Notice that 4 is the largest common divisor of 12 and 20. Similarly, 3 is a common divisor of 18 and 30, but it is not the largest, since 6 is also a common divisor. The largest common divisor of two numbers is an extremely important quantity that will frequently appear during our number theoretic excursions.

The *greatest common divisor* of two numbers a and b (not both zero) is the largest number that divides both of them. It is denoted by $\gcd(a, b)$. If $\gcd(a, b) = 1$, we say that a and b are *relatively prime*.

Two examples that we mentioned above are

$$\gcd(12, 20) = 4 \quad \text{and} \quad \gcd(18, 30) = 6.$$

Another example is

$$\gcd(225, 120) = 15.$$

We can check that this answer is correct by factoring $225 = 3^2 \cdot 5^2$ and $120 = 2^3 \cdot 3 \cdot 5$, but, in general, factoring a and b is not an efficient way to compute their greatest common divisor.¹

The most efficient method known for finding the greatest common divisors of two numbers is called the *Euclidean algorithm*. It consists of doing a sequence of divisions with remainder until the remainder is zero. We will illustrate with two examples before describing the general method.

As our first example, we will compute $\gcd(36, 132)$. The first step is to divide 132 by 36, which gives a quotient of 3 and a remainder of 24. We write this as

$$132 = 3 \times 36 + 24.$$

The next step is to take 36 and divide it by the remainder 24 from the previous step. This gives

$$36 = 1 \times 24 + 12.$$

Next we divide 24 by 12, and we find a remainder of 0,

$$24 = 2 \times 12 + 0.$$

The Euclidean algorithm says that as soon as you get a remainder of 0, the remainder from the previous step is the greatest common divisor of the original two numbers. So in this case we find that $\gcd(132, 36) = 12$.

Let's do a larger example. We will compute

$$\gcd(1160718174, 316258250).$$

Our reason for doing a large example like this is to help convince you that the Euclidean algorithm gives a far more efficient way to compute gcd's than factorization. We begin by dividing 1160718174 by 316258250, which gives 3 with a remainder of 211943424. Next we take 316258250 and divide it by 211943424. This process continues until we get a remainder of 0. The calculations are given in

¹An even less efficient way to compute the greatest common divisor of a and b is the method taught to my daughter by her fourth grade teacher, who recommended that the students make complete lists of all the divisors of a and b and then pick out the largest number that appears on both lists!

the following table:

$$\begin{array}{rcl}
 1160718174 & = & 3 \times 316258250 + 211943424 \\
 316258250 & = & 1 \times 211943424 + 104314826 \\
 211943424 & = & 2 \times 104314826 + 3313772 \\
 104314826 & = & 31 \times 3313772 + 1587894 \\
 3313772 & = & 2 \times 1587894 + 137984 \\
 1587894 & = & 11 \times 137984 + 70070 \\
 137984 & = & 1 \times 70070 + 67914 \\
 70070 & = & 1 \times 67914 + 2156 \\
 67914 & = & 31 \times 2156 + \boxed{1078} \leftarrow \text{gcd} \\
 2156 & = & 2 \times 1078 + 0
 \end{array}$$

Notice how at each step we divide a number A by a number B to get a quotient Q and a remainder R . In other words,

$$A = Q \times B + R.$$

Then at the next step we replace our old A and B with the numbers B and R and continue the process until we get a remainder of 0. At that point, the remainder R from the previous step is the greatest common divisor of our original two numbers. So the above calculation shows that

$$\text{gcd}(1160718174, 316258250) = 1078.$$

We can partly check our calculation (always a good idea) by verifying that 1078 is indeed a common divisor. Thus

$$1160718174 = 1078 \times 1076733 \quad \text{and} \quad 316258250 = 1078 \times 293375.$$

There is one more practical matter to be mentioned before we undertake a theoretical analysis of the Euclidean algorithm. If we are given A and B , how can we find the quotient Q and the remainder R ? Of course, you can always use long division, but that can be time consuming and subject to arithmetic errors if A and B are large. A pleasant alternative is to find a calculator or computer program that will automatically compute Q and R for you. However, even if you are only equipped with an inexpensive calculator, there is an easy three-step method to find Q and R .

Method to Compute Q and R on a Calculator So That $A = B \times Q + R$

1. Use the calculator to divide A by B . You get a number with decimals.
2. Discard all the digits to the right of the decimal point. This gives Q .
3. To find R , use the formula $R = A - B \times Q$.

For example, suppose that $A = 12345$ and $B = 417$. Then $A/B = 29.6043\dots$, so $Q = 29$ and $R = 12345 - 417 \cdot 29 = 252$.

We're now ready to analyze the Euclidean algorithm. The general method looks like

$$\begin{aligned} a &= q_1 \times b + r_1 \\ b &= q_2 \times r_1 + r_2 \\ r_1 &= q_3 \times r_2 + r_3 \\ r_2 &= q_4 \times r_3 + r_4 \\ &\vdots \\ r_{n-3} &= q_{n-1} \times r_{n-2} + r_{n-1} \\ r_{n-2} &= q_n \times r_{n-1} + \boxed{r_n} \leftarrow \text{gcd} \\ r_{n-1} &= q_{n+1} r_n + 0 \end{aligned}$$

If we let $r_0 = b$ and $r_{-1} = a$, then every line looks like

$$r_{i-1} = q_{i+1} \times r_i + r_{i+1}.$$

Why is the last nonzero remainder r_n a common divisor of a and b ? We start from the bottom and work our way up. The last line $r_{n-1} = q_{n+1} r_n$ shows that r_n divides r_{n-1} . Then the previous line

$$r_{n-2} = q_n \times r_{n-1} + r_n$$

shows that r_n divides r_{n-2} , since it divides both r_{n-1} and r_n . Now looking at the line above that, we already know that r_n divides both r_{n-1} and r_{n-2} , so we find that r_n also divides r_{n-3} . Moving up line by line, when we reach the second line we will already know that r_n divides r_2 and r_1 . Then the second line $b = q_2 \times r_1 + r_2$ tells us that r_n divides b . Finally, we move up to the top line and use the fact that r_n divides both r_1 and b to conclude that r_n also divides a . This completes our verification that the last nonzero remainder r_n is a common divisor of a and b .

But why is r_n the *greatest* common divisor of a and b ? Suppose that d is any common divisor of a and b . We will work our way back down the list of equations. So from the first equation $a = q_1 \times b + r_1$ and the fact that d divides both a and b , we see that d also divides r_1 . Then the second equation $b = q_2 r_1 + r_2$ shows us that d must divide r_2 . Continuing down line by line, at each stage we will know that d divides the previous two remainders r_{i-1} and r_i , and then the current line $r_{i-1} = q_{i+1} \times r_i + r_{i+1}$ will tell us that d also divides the next remainder r_{i+1} . Eventually, we reach the penultimate line $r_{n-2} = q_n \times r_{n-1} + r_n$, at which point we conclude that d divides r_n . So we have shown that if d is any common divisor of a and b then d will divide r_n . Therefore, r_n must be the greatest common divisor of a and b .

This completes our verification that the Euclidean algorithm actually computes the greatest common divisor, a fact of sufficient importance to be officially recorded.

Theorem 5.1 (Euclidean Algorithm). *To compute the greatest common divisor of two numbers a and b , let $r_{-1} = a$, let $r_0 = b$, and compute successive quotients and remainders*

$$r_{i-1} = q_{i+1} \times r_i + r_{i+1}$$

for $i = 0, 1, 2, \dots$ until some remainder r_{n+1} is 0. The last nonzero remainder r_n is then the greatest common divisor of a and b .

There remains the question of why the Euclidean algorithm always finishes. In other words, we know that the last nonzero remainder will be the desired gcd, but how do we know that we ever get a remainder that does equal 0? This is not a silly question, since it is easy to give algorithms that do not terminate; and there are even very simple algorithms for which it is not known whether or not they always terminate. Fortunately, it is easy to see that the Euclidean algorithm always terminates. The reason is simple. Each time we compute a quotient with remainder,

$$A = Q \times B + R,$$

the remainder will be between 0 and $B - 1$. This is clear, since if $R \geq B$, then we can add one more onto the quotient Q and subtract B from R . So the successive remainders in the Euclidean algorithm continually decrease:

$$b = r_0 > r_1 > r_2 > r_3 > \dots$$

But all the remainders are greater than or equal to 0, so we have a strictly decreasing sequence of nonnegative integers. Eventually, we must reach a remainder that equals 0; in fact, it is clear that we will reach a remainder of 0 in at most b steps. Fortunately, the Euclidean algorithm is far more efficient than this. You will show in the exercises that the number of steps in the Euclidean algorithm is at most seven times the *number of digits* in b . So, on a computer, it is quite feasible to compute $\text{gcd}(a, b)$ when a and b have hundreds or even thousands of digits!

Exercises

5.1. Use the Euclidean algorithm to compute each of the following gcd's.

- (a) $\text{gcd}(12345, 67890)$ (b) $\text{gcd}(54321, 9876)$

5.2.  Write a program to compute the greatest common divisor $\text{gcd}(a, b)$ of two integers a and b . Your program should work even if one of a or b is zero. Make sure that you don't go into an infinite loop if a and b are both zero!

5.3. Let $b = r_0, r_1, r_2, \dots$ be the successive remainders in the Euclidean algorithm applied to a and b . Show that after every two steps, the remainder is reduced by at least one half. In other words, verify that

$$r_{i+2} < \frac{1}{2}r_i \quad \text{for every } i = 0, 1, 2, \dots$$

Conclude that the Euclidean algorithm terminates in at most $2 \log_2(b)$ steps, where \log_2 is the logarithm to the base 2. In particular, show that the number of steps is at most seven times the number of digits in b . [Hint. What is the value of $\log_2(10)$?]

5.4. A number L is called a common multiple of m and n if both m and n divide L . The smallest such L is called the *least common multiple of m and n* and is denoted by $\text{LCM}(m, n)$. For example, $\text{LCM}(3, 7) = 21$ and $\text{LCM}(12, 66) = 132$.

- (a) Find the following least common multiples.
 (i) $\text{LCM}(8, 12)$ (ii) $\text{LCM}(20, 30)$ (iii) $\text{LCM}(51, 68)$ (iv) $\text{LCM}(23, 18)$.
 (b) For each of the LCMs that you computed in (a), compare the value of $\text{LCM}(m, n)$ to the values of m , n , and $\text{gcd}(m, n)$. Try to find a relationship.
 (c) Give an argument proving that the relationship you found is correct for all m and n .
 (d) Use your result in (b) to compute $\text{LCM}(301337, 307829)$.
 (e) Suppose that $\text{gcd}(m, n) = 18$ and $\text{LCM}(m, n) = 720$. Find m and n . Is there more than one possibility? If so, find all of them.

5.5. The “ $3n + 1$ algorithm” works as follows. Start with any number n . If n is even, divide it by 2. If n is odd, replace it with $3n + 1$. Repeat. So, for example, if we start with 5, we get the list of numbers

$$5, 16, 8, 4, 2, 1, 4, 2, 1, 4, 2, 1, \dots,$$

and if we start with 7, we get

$$7, 22, 11, 34, 17, 52, 26, 13, 40, 20, 10, 5, 16, 8, 4, 2, 1, 4, 2, 1, \dots$$

Notice that if we ever get to 1 the list just continues to repeat with 4, 2, 1's. In general, one of the following two possibilities will occur:²

- (i) We may end up repeating some number a that appeared earlier in our list, in which case the block of numbers between the two a 's will repeat indefinitely. In this case we say that the algorithm *terminates* at the last nonrepeated value, and the number of distinct entries in the list is called the *length of the algorithm*. For example, the algorithm terminates at 1 for both 5 and 7. The length of the algorithm for 5 is 6, and the length of the algorithm for 7 is 17.
 (ii) We may never repeat the same number, in which case we say that the algorithm does not terminate.

²There is, of course, a third possibility. We may get tired of computing and just stop working, in which case one might say that the algorithm terminates due to exhaustion of the computer!

- (a) Find the length and terminating value of the $3n+1$ algorithm for each of the following starting values of n :

$$(i) n = 21 \quad (ii) n = 13 \quad (iii) n = 31$$

- (b) Do some further experimentation and try to decide whether the $3n + 1$ algorithm always terminates and, if so, at what value(s) it terminates.
- (c) Assuming that the algorithm terminates at 1, let $L(n)$ be the length of the algorithm for starting value n . For example, $L(5) = 6$ and $L(7) = 17$. Show that if $n = 8k + 4$ with $k \geq 1$, then $L(n) = L(n + 1)$. [*Hint*. What does the algorithm do to the starting values $8k + 4$ and $8k + 5$?]
- (d) Show that if $n = 128k + 28$ then $L(n) = L(n + 1) = L(n + 2)$.
- (e) Find some other conditions, similar to those in (c) and (d), for which consecutive values of n have the same length. (It might be helpful to begin by using the next exercise to accumulate some data.)

5.6.  Write a program to implement the $3n + 1$ algorithm described in the previous exercise. The user will input n and your program should return the length $L(n)$ and the terminating value $T(n)$ of the $3n + 1$ algorithm. Use your program to create a table giving the length and terminating value for all starting values $1 \leq n \leq 100$.

Chapter 6

Linear Equations and the Greatest Common Divisor

Given two whole numbers a and b , we are going to look at all the possible numbers we can get by adding a multiple of a to a multiple of b . In other words, we will consider all numbers obtained from the formula

$$ax + by$$

when we substitute all possible integers for x and y . Note that we are going to allow both positive and negative values for x and y . For example, we could take $a = 42$ and $b = 30$. Some of the values of $ax + by$ for this a and b are given in the following table:

	$x = -3$	$x = -2$	$x = -1$	$x = 0$	$x = 1$	$x = 2$	$x = 3$
$y = -3$	-216	-174	-132	-90	-48	-6	36
$y = -2$	-186	-144	-102	-60	-18	24	66
$y = -1$	-156	-114	-72	-30	12	54	96
$y = 0$	-126	-84	-42	0	42	84	126
$y = 1$	-96	-54	-12	30	72	114	156
$y = 2$	-66	-24	18	60	102	144	186
$y = 3$	-36	6	48	90	132	174	216

Table of Values of $42x + 30y$

Our first observation is that every entry in the table is divisible by 6. This is not surprising, since both 42 and 30 are divisible by 6, so every number of the form $42x + 30y = 6(7x + 5y)$ is a multiple of 6. More generally, it is clear that every number of the form $ax + by$ is divisible by $\gcd(a, b)$, since both a and b are divisible by $\gcd(a, b)$.

A second observation, which is somewhat more surprising, is that the greatest common divisor of 42 and 30, which is 6, actually appears in our table. Thus from the table we see that

$$42 \cdot (-2) + 30 \cdot 3 = 6 = \gcd(42, 30).$$

Further examples suggest the following conclusion:

The smallest positive value of
 $ax + by$
 is equal to $\gcd(a, b)$.

There are many ways to prove that this is true. We will take a constructive approach, via the Euclidean algorithm, which has the advantage of giving a procedure for finding the appropriate values of x and y . In other words, we are going to describe a method of finding integers x and y that are solutions to the equation

$$ax + by = \gcd(a, b).$$

Since, as we have already observed, every number $ax + by$ is divisible by $\gcd(a, b)$, it will follow that the smallest positive value of $ax + by$ is precisely $\gcd(a, b)$.

How might we solve the equation $ax + by = \gcd(a, b)$? If a and b are small, we might be able to guess a solution. For example, the equation

$$10x + 35y = 5$$

has the solution $x = -3$ and $y = 1$, and the equation

$$7x + 11y = 1$$

has the solution $x = -3$ and $y = 2$. We also notice that there can be more than one solution, since $x = 8$ and $y = -5$ is also a solution to $7x + 11y = 1$.

However, if a and b are large, neither guesswork nor trial and error is going to be helpful. We are going to start by illustrating the Euclidean algorithm method for solving $ax + by = \gcd(a, b)$ with a particular example. So we are going to try to solve

$$22x + 60y = \gcd(22, 60).$$

The first step is to perform the Euclidean algorithm to compute the gcd. We find

$$\begin{aligned} 60 &= 2 \times 22 + 16 \\ 22 &= 1 \times 16 + 6 \\ 16 &= 2 \times 6 + 4 \\ 6 &= 1 \times 4 + 2 \\ 4 &= 2 \times 2 + 0 \end{aligned}$$

This shows that $\gcd(22, 60) = 2$, a fact that is clear without recourse to the Euclidean algorithm. However, the Euclidean algorithm computation is important because we're going to use the intermediate quotients and remainders to solve the equation $22x + 60y = 2$. The first step is to rewrite the first equation as

$$16 = a - 2b, \quad \text{where we let } a = 60 \text{ and } b = 22.$$

We next substitute this value into the 16 appearing in the second equation. This gives (remember that $b = 22$)

$$b = 1 \times 16 + 6 = 1 \times (a - 2b) + 6.$$

Rearranging this equation to isolate the remainder 6 yields

$$6 = b - (a - 2b) = -a + 3b.$$

Now substitute the values 16 and 6 into the next equation, $16 = 2 \times 6 + 4$:

$$a - 2b = 16 = 2 \times 6 + 4 = 2(-a + 3b) + 4.$$

Again we isolate the remainder 4, yielding

$$4 = (a - 2b) - 2(-a + 3b) = 3a - 8b.$$

Finally, we use the equation $6 = 1 \times 4 + 2$ to get

$$-a + 3b = 6 = 1 \times 4 + 2 = 1 \times (3a - 8b) + 2.$$

Rearranging this equation gives the desired solution

$$-4a + 11b = 2.$$

(We should check our solution: $-4 \times 60 + 11 \times 22 = -240 + 242 = 2$.)

We can summarize the above computation in the following efficient tabular form. Note that the left-hand equations are the Euclidean algorithm, and the right-hand equations compute the solution to $ax + by = \gcd(a, b)$.

$$\begin{array}{l|l}
 a = 2 \times b + 16 & 16 = a - 2b \\
 b = 1 \times 16 + 6 & 6 = b - 1 \times 16 \\
 & = b - 1 \times (a - 2b) \\
 & = -a + 3b \\
 16 = 2 \times 6 + 4 & 4 = 16 - 2 \times 6 \\
 & = (a - 2b) - 2 \times (-a + 3b) \\
 & = 3a - 8b \\
 6 = 1 \times 4 + 2 & 2 = 6 - 1 \times 4 \\
 & = (-a + 3b) - 1 \times (3a - 8b) \\
 & = -4a + 11b \\
 4 = 2 \times 2 + 0 &
 \end{array}$$

Why does this method work? As the following table makes clear, we start with the first two lines of the Euclidean algorithm, which involve the quantities a and b , and work our way down.

$$\begin{array}{l|l}
 a = q_1b + r_1 & r_1 = a - q_1b \\
 b = q_2r_1 + r_2 & r_2 = b - q_2r_1 \\
 & = b - q_2(a - q_1b) \\
 & = -q_2a + (1 + q_1q_2)b \\
 r_1 = q_3r_2 + r_3 & r_3 = r_1 - q_3r_2 \\
 & = (a - q_1b) - q_3(-q_2a + (1 + q_1q_2)b) \\
 & = (1 + q_2q_3)a - (q_1 + q_3 + q_1q_2q_3)b \\
 \vdots & \vdots
 \end{array}$$

As we move from line to line, we will continually be forming equations that look like

$$\text{latest remainder} = \text{some multiple of } a \text{ plus some multiple of } b.$$

Eventually, we get down to the last nonzero remainder, which we know is equal to $\gcd(a, b)$, and this gives the desired solution to the equation $\gcd(a, b) = ax + by$.

A larger example with $a = 12453$ and $b = 2347$ is given in tabular form on top of the next page. As before, the left-hand side is the Euclidean algorithm and the right-hand side solves $ax + by = \gcd(a, b)$. We see that $\gcd(12453, 2347) = 1$ and that the equation $12453x + 2347y = 1$ has the solution $(x, y) = (304, -1613)$.

We now know that the equation

$$ax + by = \gcd(a, b)$$

always has a solution in integers x and y . The final topic we discuss in this section is the question of how many solutions it has, and how to describe all the solutions. Let's start with the case that a and b are relatively prime, that is, $\gcd(a, b) = 1$, and suppose that (x_1, y_1) is a solution to the equation

$$ax + by = 1.$$

We can create additional solutions by subtracting a multiple of b from x_1 and adding the same multiple of a onto y_1 . In other words, for any integer k we obtain a new solution $(x_1 + kb, y_1 - ka)$.¹ We can check that this is indeed a solution by computing

$$a(x_1 + kb) + b(y_1 - ka) = ax_1 + akb + by_1 - bka = ax_1 + by_1 = 1.$$

¹Geometrically, we are starting from the known point (x_1, y_1) on the line $ax + by = 1$ and using the fact that the line has slope $-a/b$ to find new points $(x_1 + t, y_1 - (a/b)t)$. To get new points with integer coordinates, we need to let t be a multiple of b . Substituting $t = kb$ gives the new integer solution $(x_1 + kb, y_1 - ka)$.

$a = 5 \times b + 718$	$718 = a - 5b$
$b = 3 \times 718 + 193$	$193 = b - 3 \times 718$
	$= b - 3 \times (a - 5b)$
	$= -3a + 16b$
$718 = 3 \times 193 + 139$	$139 = 718 - 3 \times 193$
	$= (a - 5b) - 3 \times (-3a + 16b)$
	$= 10a - 53b$
$193 = 1 \times 139 + 54$	$54 = 193 - 139$
	$= (-3a + 16b) - (10a - 53b)$
	$= -13a + 69b$
$139 = 2 \times 54 + 31$	$31 = 139 - 2 \times 54$
	$= (10a - 53b) - 2 \times (-13a + 69b)$
	$= 36a - 191b$
$54 = 1 \times 31 + 23$	$23 = 54 - 31$
	$= -13a + 69b - (36a - 191b)$
	$= -49a + 260b$
$31 = 1 \times 23 + 8$	$8 = 31 - 23$
	$= 36a - 191b - (-49a + 260b)$
	$= 85a - 451b$
$23 = 2 \times 8 + 7$	$7 = 23 - 2 \times 8$
	$= (-49a + 260b) - 2 \times (85a - 451b)$
	$= -219a + 1162b$
$8 = 1 \times 7 + 1$	$1 = 8 - 7$
	$= 85a - 451b - (-219a + 1162b)$
	$= 304a - 1613b$
$7 = 7 \times 1 + 0$	

So, for example, if we start with the solution $(-1, 2)$ to $5x + 3y = 1$, we obtain new solutions $(-1 + 3k, 2 - 5k)$. Note that the integer k is allowed to be positive, negative, or zero. Putting in particular values of k gives the solutions

$$\dots (-13, 22), (-10, 17), (-7, 12), (-4, 7), (-1, 2), \\ (2, -3), (5, -8), (8, -13), (11, -18) \dots$$

Still looking at the case that $\gcd(a, b) = 1$, we can show that this procedure gives all possible solutions. Suppose that we are given two solutions (x_1, y_1) and (x_2, y_2) to the equation $ax + by = 1$. In other words,

$$ax_1 + by_1 = 1 \quad \text{and} \quad ax_2 + by_2 = 1.$$

We are going to multiply the first equation by y_2 , multiply the second equation by y_1 , and subtract. This will eliminate b and, after a little bit of algebra, we are

left with

$$ax_1y_2 - ax_2y_1 = y_2 - y_1.$$

Similarly, if we multiply the first equation by x_2 , multiply the second equation by x_1 , and subtract, we find that

$$bx_2y_1 - bx_1y_2 = x_2 - x_1.$$

So if we let $k = x_2y_1 - x_1y_2$, then we find that

$$x_2 = x_1 + kb \quad \text{and} \quad y_2 = y_1 - ka.$$

This means that the second solution (x_2, y_2) is obtained from the first solution (x_1, y_1) by adding a multiple of b onto x_1 and subtracting the same multiple of a from y_1 . So every solution to $ax + by = 1$ can be obtained from the initial solution (x_1, y_1) by substituting different values of k into $(x_1 + kb, y_1 - ka)$.

What happens if $\gcd(a, b) > 1$? To make the formulas look a little bit simpler, we will let $g = \gcd(a, b)$. We know from the Euclidean algorithm method that there is at least one solution (x_1, y_1) to the equation

$$ax + by = g.$$

But g divides both a and b , so (x_1, y_1) is a solution to the simpler equation

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

Now our earlier work applies, so we know that every other solution can be obtained by substituting values for k in the formula

$$\left(x_1 + k \cdot \frac{b}{g}, y_1 - k \cdot \frac{a}{g} \right).$$

This completes our description of the solutions to the equation $ax + by = g$, as summarized in the following theorem.

Theorem 6.1 (Linear Equation Theorem). *Let a and b be nonzero integers, and let $g = \gcd(a, b)$. The equation*

$$ax + by = g$$

always has a solution (x_1, y_1) in integers, and this solution can be found by the Euclidean algorithm method described earlier. Then every solution to the equation can be obtained by substituting integers k into the formula

$$\left(x_1 + k \cdot \frac{b}{g}, y_1 - k \cdot \frac{a}{g} \right).$$

For example, we saw that the equation

$$60x + 22y = \gcd(60, 22) = 2$$

has the solution $x = -4$, $y = 11$. Then our Linear Equation Theorem says that every solution is obtained from the formula

$$(-4 + 11k, 11 - 30k) \quad \text{with } k \text{ any integer.}$$

In particular, if we want a solution with x positive, then we can take $k = 1$, which gives the smallest such solution $(x, y) = (7, -19)$.

In this chapter we have shown that the equation

$$ax + by = \gcd(a, b)$$

always has a solution. This fact is extremely important for both theoretical and practical reasons, and we will be using it repeatedly in our number theoretic investigations. For example, we will need to solve the equation $ax + by = 1$ when we study cryptography in Chapter 18. And in the next chapter we will use this equation for our theoretical study of factorization of numbers into primes.

Exercises

6.1. (a) Find a solution in integers to the equation

$$12345x + 67890y = \gcd(12345, 67890).$$

(b) Find a solution in integers to the equation

$$54321x + 9876y = \gcd(54321, 9876).$$

6.2. Describe all integer solutions to each of the following equations.

(a) $105x + 121y = 1$

(b) $12345x + 67890y = \gcd(12345, 67890)$

(c) $54321x + 9876y = \gcd(54321, 9876)$

6.3.  The method for solving $ax + by = \gcd(a, b)$ described in this chapter involves a considerable amount of manipulation and back substitution. This exercise describes an alternative way to compute x and y that is especially easy to implement on a computer.

(a) Show that the algorithm described in Figure 6.1 computes the greatest common divisor g of the positive integers a and b , together with a solution (x, y) in integers to the equation $ax + by = \gcd(a, b)$.

(b) Implement the algorithm on a computer using the computer language of your choice.

- (c) Use your program to compute $g = \gcd(a, b)$ and integer solutions to $ax + by = g$ for the following pairs (a, b) .
- (i) (19789, 23548) (ii) (31875, 8387) (iii) (22241739, 19848039)
- (d) What happens to your program if $b = 0$? Fix the program so that it deals with this case correctly.
- (e) For later applications it is useful to have a solution with $x > 0$. Modify your program so that it always returns a solution with $x > 0$. [Hint. If (x, y) is a solution, then so is $(x + b, y - a)$.]

- (1) Set $x = 1$, $g = a$, $v = 0$, and $w = b$.
- (2) If $w = 0$ then set $y = (g - ax)/b$ and return the values (g, x, y) .
- (3) Divide g by w with remainder, $g = qw + t$, with $0 \leq t < w$.
- (4) Set $s = x - qv$.
- (5) Set $(x, g) = (v, w)$.
- (6) Set $(v, w) = (s, t)$.
- (7) Go to Step (2).

Figure 6.1: Efficient algorithm to solve $ax + by = \gcd(a, b)$

- 6.4. (a)** Find integers x , y , and z that satisfy the equation

$$6x + 15y + 20z = 1.$$

- (b) Under what conditions on a, b, c is it true that the equation

$$ax + by + cz = 1$$

has a solution? Describe a general method of finding a solution when one exists.

- (c) Use your method from (b) to find a solution in integers to the equation

$$155x + 341y + 385z = 1.$$

- 6.5.** Suppose that $\gcd(a, b) = 1$. Prove that for every integer c , the equation $ax + by = c$ has a solution in integers x and y . [Hint. Find a solution to $au + bv = 1$ and multiply by c .] Find a solution to $37x + 47y = 103$. Try to make x and y as small as possible.

- 6.6.** Sometimes we are only interested in solutions to $ax + by = c$ using nonnegative values for x and y .

- (a) Explain why the equation $3x + 5y = 4$ has no solutions with $x \geq 0$ and $y \geq 0$.
- (b) Make a list of some of the numbers of the form $3x + 5y$ with $x \geq 0$ and $y \geq 0$. Make a conjecture as to which values are not possible. Then prove that your conjecture is correct.

- (c) For each of the following values of (a, b) , find the largest number that is not of the form $ax + by$ with $x \geq 0$ and $y \geq 0$.
- (i) $(a, b) = (3, 7)$ (ii) $(a, b) = (5, 7)$ (iii) $(a, b) = (4, 11)$.
- (d) Let $\gcd(a, b) = 1$. Using your results from (c), find a conjectural formula in terms of a and b for the largest number that is not of the form $ax + by$ with $x \geq 0$ and $y \geq 0$? Check your conjecture for at least two more values of (a, b) .
- (e) Prove that your conjectural formula in (d) is correct.
- (f) Try to generalize this problem to sums of three terms $ax + by + cz$ with $x \geq 0$, $y \geq 0$, and $z \geq 0$. For example, what is the largest number that is not of the form $6x + 10y + 15z$ with nonnegative x, y, z ?