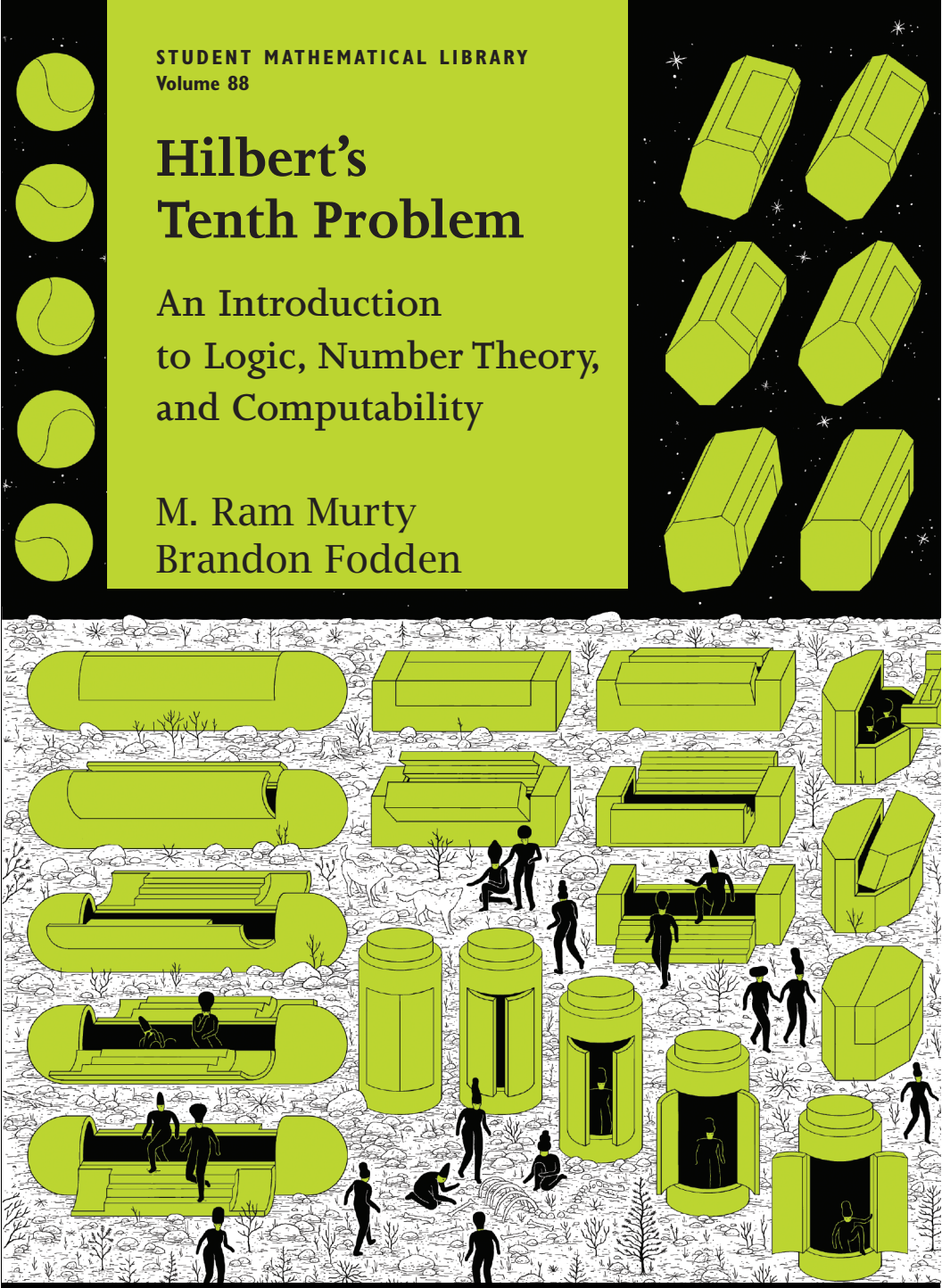
STUDENT MATHEMATICAL LIBRARY
Volume 88

Hilbert's Tenth Problem

An Introduction to Logic, Number Theory, and Computability

M. Ram Murty
Brandon Fodden



AMS American Mathematical Society

Hilbert's Tenth Problem

An Introduction to Logic, Number Theory, and Computability

STUDENT MATHEMATICAL LIBRARY
Volume 88

Hilbert's Tenth Problem

An Introduction to Logic, Number Theory, and Computability

M. Ram Murty
Brandon Fodden

American Mathematical Society
Providence, Rhode Island

2010 *Mathematics Subject Classification.* Primary 11U05, 12L05.

For additional information and updates on this book, visit
www.ams.org/bookpages/stml-88

Library of Congress Cataloging-in-Publication Data

Names: Murty, Maruti Ram, author. | Fodden, Brandon, 1979– author.
Title: Hilbert's tenth problem : an introduction to logic, number theory, and computability / M. Ram Murty, Brandon Fodden.
Description: Providence, Rhode Island : American Mathematical Society, [2019] | Series: Student mathematical library ; volume 88 | Includes bibliographical references and index.
Identifiers: LCCN 2018061472 | ISBN 9781470443993 (alk. paper)
Subjects: LCSH: Hilbert's tenth problem. | Number theory–Problems, exercises, etc. | Mathematical recreations–Problems, exercises, etc. | Hilbert, David, 1862–1943. | AMS: Number theory – Connections with logic – Decidability. msc | Field theory and polynomials – Connections with logic – Decidability. msc
Classification: LCC QA242 .M8945 2019 | DDC 512.7/4–dc23
LC record available at https://lccn.loc.gov/2018061472

10 9 8 7 6 5 4 3 2 1 24 23 22 21 20 19

An algebra of mind, a scheme of sense,
A symbol language without depth or wings,
A power to handle deftly outward things
Are our scant earnings of intelligence.
The Truth is greater and asks deeper ways.

- Sri Aurobindo, "Discoveries of Science II" in *Collected Poems*

Contents

Preface

In 1980, the senior author (MRM) had the grand privilege of meeting Sarvadaman Chowla at the Institute for Advanced Study in Princeton, New Jersey. Chowla had written a small book titled *The Riemann Hypothesis and Hilbert's Tenth Problem* in 1965 and so this was an opportunity to ask him about the seemingly strange title and how it came to be. Was there a connection between the two? Chowla replied, "I don't know. These two problems have always fascinated me and so I chose that as the title." He went on to say that the book was largely an inspired work, written in a single night, and it represents his selection of beautiful pearls from number theory. It was not meant to be a textbook but more of an invitation for further study and "to stimulate the reader".

But the fact of the matter is that the two problems *are* related as we discovered only much later in the work of Martin Davis, Hilary Putnam, Julia Robinson and Yuri Matiyasevič. In fact, many of the famous Hilbert problems are interconnected. This interconnectedness can be used as the focus for mathematical instruction. And it can be done with very few prerequisites. This is the *raison d'être* of this book.

Some of the Hilbert problems such as the Riemann hypothesis (the eighth problem) are still open. The others that have been solved required formidable background and preparation. Hilbert's tenth

problem is different in that a basic introduction to elementary number theory and mathematical logic suffices to understand the proof, and this can be done in a relatively short time. In addition to the grand arrangement of mathematical ideas, Hilbert's tenth problem has a colourful cast of characters, many of them tragic heroes, who pondered deeply regarding the enigma of the human being and the nature of mathematical truth.

Hilbert's tenth problem and its solution represent in microcosm the riddle of human life itself and its meaning. This mélange of philosophical and mathematical conundrums are the mysteries that confront us. In many ways, this book is not meant to be a textbook, but rather an invitation to explore further. As Chowla would say, we hope "to stimulate the reader".

M. Ram Murty and Brandon Fodden
Kingston and Ottawa, Ontario
July 2018

Acknowledgments

This book is based on an upper-level undergraduate course given at Queen's University in Ontario in the winter semester of 2007 by the senior author (MRM). The class consisted of primarily undergraduates, several graduate students, and a few post-doctoral fellows. There were also students from the philosophy department. Given the diverse backgrounds of the students, the mathematical prerequisites were kept to a bare minimum requiring only familiarity with basic calculus and linear algebra. The course covered the contents of the first five chapters by first introducing students to logical notation, then elementary number theory, and gradually to notions of computability and decidability and, finally, the proof of Hilbert's tenth problem. The last two chapters were added later and were culled from graduate seminars conducted since the time the course was first given. They require more advanced background, especially the last chapter. If the student is willing to take some of the background material in those chapters on faith, they will acquire a panoramic view of some recent discoveries and new directions. We feel that this assemblage of subject matter can make an excellent introduction to this fascinating topic and can take the student to the frontiers of current research. We thank Kumar Murty, Hector Pasten, and the referees for their comments on an earlier version of this book. We are grateful to Ina Mette and the American Mathematical Society for taking interest in publishing this book, and to Marcia Almeida at the AMS for much help with preparing the manuscript.

Introduction

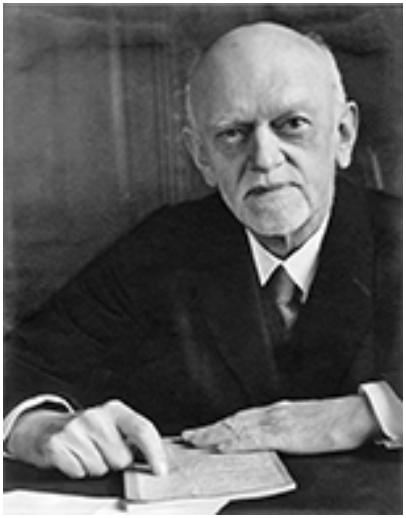
In 1900, at the second International Congress of Mathematicians (ICM), taking place in Paris, David Hilbert (1862–1943) presented a list of twenty-three problems that he felt were fundamentally important, and would influence the direction of mathematics in the 20th century.[1] In his own words (in translation):[2]

> The supply of problems in mathematics is inexhaustible, and as soon as one problem is solved numerous others come forth in its place. Permit me in the following, tentatively as it were, to mention particular definite problems, drawn from various branches of mathematics, from the discussion of which an advancement of science may be expected.

In his tenth problem, Hilbert asked for an algorithm that, when given an arbitrary Diophantine equation, will determine whether the equation has integer solutions or not. The solution to this problem, that there is no such algorithm, is one of the remarkable achievements of 20th-century mathematics. When such dramatic advances in a

[1] The address was given in German, and he presented ten of the problems. In a paper written in French appearing in the proceedings of the congress, he included the full list of twenty-three problems.

[2] Hilbert's address was translated and published in the *Bulletin of the American Mathematical Society*; see [**Hil02**].



DAVID HILBERT (Photo courtesy the Archives of the Mathematisches Forschungsinstitut Oberwolfach)

discipline are made, it is rare that one can explain the solution in simple terms to the nonexpert. Fortunately, this is not the case with Hilbert's tenth problem. All that is needed to understand how the solution has been put together is an elementary knowledge of the rudiments of logic and elementary number theory, both topics being at a level accessible to an undergraduate student of mathematics. It is the purpose of this monograph to explain this work in as simple a language as possible. In fact, we feel it is a splendid way to introduce the student to both logic and number theory through such a motivated introduction with Hilbert's tenth problem as the focus.

That Hilbert's tenth problem has a negative solution, in that the algorithm that Hilbert sought does not exist, might have surprised Hilbert. Still, he expected that such things may occur. In his 1900 address, he commented:

> Occasionally it happens that we seek the solution under insufficient hypotheses or in an incorrect sense, and for this reason do not succeed. The problem then arises: to show the impossibility of

> the solution under the given hypotheses, or in the sense contemplated. Such proofs of impossibility were effected by the ancients, for instance when they showed that the ratio of the hypotenuse to the side of an isosceles right triangle is irrational. In later mathematics, the question as to the impossibility of certain solutions plays a pre-eminent part, and we perceive in this way that old and difficult problems, such as the proof of the axiom of parallels, the squaring of the circle, or the solution of equations of the fifth degree by radicals, have finally found fully satisfactory and rigorous solutions, although in another sense than that originally intended. It is probably this important fact along with other philosophical reasons that gives rise to the conviction (which every mathematician shares, but which no one has as yet supported by a proof) that every definite mathematical problem must necessarily be susceptible of an exact settlement, either in the form of an actual answer to the question asked or by the proof of the impossibility of its solution and therewith the necessary failure of all attempts.

In this book, we touch on several of Hilbert's problems. His first problem, the continuum hypothesis, is discussed in Sections 2.3 and 4.3. In his second problem, Hilbert asked for a proof of the consistency of the axioms of arithmetic. We discuss Kurt Gödel's momentous result on this problem in Section 4.4. We briefly mention Hilbert's seventh problem, on the transcendence of a^b when $a \neq 0, 1$ is algebraic and b is irrational and algebraic, in Section 1.2. Hilbert's eighth problem includes the Riemann hypothesis, Goldbach's conjecture, and the twin prime conjecture, which are discussed in Section 6.3.

The scope of this book is broad. A self-contained rigorous treatment of all the topics covered by this book would more than triple its length. Our hope is to introduce the reader to numerous topics in

logic, number theory, and computability. The interested reader can then undertake further study in these areas. To that end, a list of references for further reading is presented at the end of most chapters.

The first four chapters develop the rudimentary notions of set theory, elementary number theory, and logic needed for a complete self-contained proof of Hilbert's tenth problem in Chapter 5. Some applications of the solution to Hilbert's tenth problem are covered in Chapter 6. This material is accessible to the undergraduate student and can be covered in a semester course.[3] The final chapter aims to introduce the aspiring student to current research on this topic, namely Hilbert's tenth problem over number fields. This chapter requires more mathematical maturity and is intended for the advanced student. Undoubtedly, there is more research to be done and it is our hope that the reader is thus taken to the frontier of the existing knowledge on this topic so that he or she may survey what is known and what is unknown.

If one is just looking to understand the solution to Hilbert's tenth problem, which is given in Chapter 5, then Sections 3.4 and 4.2 are necessary, provided one is comfortable using an informal definition of an algorithm. A more careful discussion of algorithms and computability is given in Section 4.1. In Chapter 6, some applications of the solution to Hilbert's tenth problem are given. These use the material developed in Chapter 2 and Sections 4.3 and 4.4. However, we feel that by reading through the entire book, one will get a better sense of the interplay between the topics covered by this book and an understanding of some of the most important problems in logic and number theory of the past 150 years.

[3] An appendix containing preliminary material is included.

Chapter 1

Cantor and Infinity

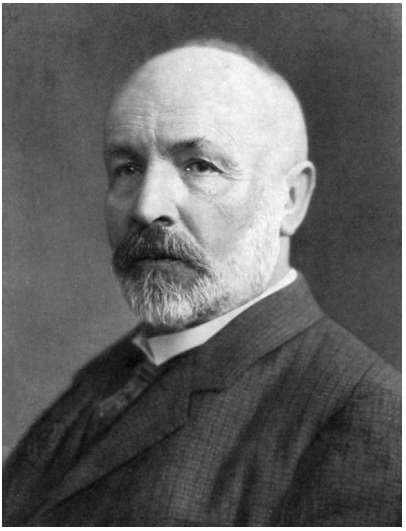
1.1. Countable Sets

There are some philosophers that deny the existence of infinity. These are the "finitists". They argue that since we have never seen an infinite collection of things, infinity does not exist. When presented with notions of mathematics that lead one to the idea of infinity, they argue that it may be that the universe is working modulo p for some large prime p!

Georg Cantor (1845–1918) can be said to be the founder of set theory and, more generally, the mathematical theory of infinite numbers. He was born in St. Petersburg into a merchant family that settled in Germany in 1856. He studied in Zürich and then Berlin where he obtained his degree in 1867. In 1869 he became a lecturer at the University of Halle in Germany and served there as a professor from 1879 to 1913. He put forth the *continuum hypothesis* (which will be described at the end of this chapter) and attempted to solve it. Perhaps under the strain of these efforts as well as initial opposition to his new ideas concerning infinity, he suffered from depression which may have eventually contributed to his death. The celebrated physicist, Stephen Hawking [**Haw**] wrote, "Georg Cantor scaled the

peaks of infinity and then plunged into the deepest abysses of the mind: mental depression."[1]

Cantor's doctoral thesis was in number theory. Later, he introduced the concepts of *ordinal numbers* and *cardinal numbers*, which we discuss in Chapter 2. Using this theory, he proved a number of results that compare the sizes of infinite sets, many of which are given here.



Georg Cantor (Photo source: Wikipedia)

A set S is said to be *countably infinite* if it can be put in one-to-one correspondence with the natural numbers (that is, if there is a bijection between $\mathbb{N} = \{0, 1, 2, \ldots\}$ and S). A *countable* set is either finite or countably infinite. If a set is not countable, it is called *uncountable*. Since the function that sends n to $2n$ is a

[1]Hawking edited the book *God Created the Integers* in which he penned short essays on about two dozen mathematical giants, with Cantor being one of them. Unfortunately, these essays were not copy edited properly and there is a serious error on page 1132. Responding to a question of Dedekind as to whether an infinite set can be defined without referring to the natural numbers, Hawking tries to give Cantor's reply. Thus, the first sentence of the last paragraph on page 1132 should be: "Cantor answered his first question by defining a set as being infinite if it could be put into a one to one correspondence with a proper subset of itself."

bijection between $\mathbb{N}$ and the set of even natural numbers, the set of even natural numbers is countably infinite. The set of integers, denoted by $\mathbb{Z}$, is also countably infinite since we may define a map $f : \mathbb{N} \to \mathbb{Z}$ by setting $f(0) = 0$, $f(1) = 1$, $f(2) = -1$, $f(3) = 2$, $f(4) = -2$, $f(5) = 3$, $f(6) = -3$, and so on.

n	0	1	2	3	4	5	6	7	8	9	10	$\cdots$
$f(n)$	0	1	-1	2	-2	3	-3	4	-4	5	-5	$\cdots$

More generally, we set

$$f(n) = (-1)^{n+1} \left\lfloor \frac{n+1}{2} \right\rfloor,$$

where $\lfloor x \rfloor$ is the floor function, which returns the greatest integer less than or equal to x. This function is easily verified to be injective and surjective.

What about $\mathbb{Q}$, the set of rational numbers? Could it be that $\mathbb{Q}$ is countably infinite? Since any positive rational number can be written as a/b for some natural numbers a and b, we are led to consider the problem of determining if the set $\mathbb{N} \times \mathbb{N}$ of ordered pairs of natural numbers is countably infinite. That is, is there a one-to-one correspondence between $\mathbb{N} \times \mathbb{N}$ and $\mathbb{N}$? Cantor discovered an explicit function, given by

$$P(x, y) = \frac{(x+y)(x+y+1)}{2} + x,$$

that sets up a one-to-one correspondence between $\mathbb{N} \times \mathbb{N}$ and $\mathbb{N}$. This function is called Cantor's pairing function. A table for the first few values of $P(x, y)$ is given below.

		y					
		0	1	2	3	4	5
x	0	0	1	3	6	10	15
	1	2	4	7	11	16	22
	2	5	8	12	17	23	30
	3	9	13	18	24	31	39
	4	14	19	25	32	40	49
	5	20	26	33	41	50	60

Cantor found the pairing function via a diagonal method of enumeration. That is, he began his list of pairs as

$$(0,0),(0,1),(1,0),(0,2),(1,1),(2,0),(0,3),(1,2),(2,1),(3,0),\ldots.$$

Let us note that we can group the pairs (a,b) according to the sum $a+b$. There are only finitely many such pairs in any group. Corresponding to the sum k, we see that there are $k+1$ such pairs. Now given an ordered pair (x,y), the group it lies in is determined by $k=x+y$. Before we reach this group, the number of ordered pairs we encounter is

$$1+2+\cdots+k=\frac{k(k+1)}{2}.$$

Having reached the group with sum k, to reach (x,y) we have to proceed through

$$(0,k),(1,k-1),\ldots,(x,y),$$

which encompass $x+1$ additional pairs. Thus the pair (x,y) is in the

$$\frac{k(k+1)}{2}+x+1=\frac{(x+y)(x+y+1)}{2}+x+1$$

position in the listing. Since we want the first listed pair to be mapped to 0, the second to be mapped to 1, and so on, subtracting 1 yields the pairing function $P(x,y)$.[2]

Using the above functions f and P, it follows that $\mathbb{Z}\times\mathbb{Z}$ is also countably infinite, for one may show that $h:\mathbb{Z}\times\mathbb{Z}\to\mathbb{N}$ given by $h(x,y)=P(f^{-1}(x),f^{-1}(y))$ is bijective. By a similar argument, $A\times B$ is also countably infinite for countably infinite A and B. By iteration, it follows that $\mathbb{Z}^n$, the set of all ordered n-tuples of elements of $\mathbb{Z}$, is countably infinite for any positive natural number n, as is $\mathbb{N}^n$.

Since Cantor's pairing function $P(x,y)$ is bijective, $F=P^{-1}$ is a bijective map from $\mathbb{N}$ to $\mathbb{N}\times\mathbb{N}$. To the ordered pair (a,b) we may associate the positive rational number $(a+1)/(b+1)$, and thus use F to list the positive rational numbers, agreeing to skip any previously

[2]In this context, we mention a result of Rudolf Fueter (1880–1950) and George Pólya (1887–1985). It is an open question to determine all bijective polynomial maps between $\mathbb{N}\times\mathbb{N}$ and $\mathbb{N}$. Fueter and Pólya showed that if we restrict our attention to quadratic polynomials, then essentially Cantor's pairing function (up to permutation) is the only one. Their proof uses the transcendence of π, as proved by Ferdinand von Lindemann (1852–1939), as well as nontrivial analytic number theory regarding error terms in lattice point enumerations.

listed numbers. This listing yields a bijection between $\mathbb{N}$ and the positive rational numbers. In particular, since

$$F(0) = (0,0),\ F(1) = (0,1),\ F(2) = (1,0),\ F(3) = (0,2),$$
$$F(4) = (1,1),\ F(5) = (2,0),\ \ldots,$$

our bijection g between $\mathbb{N}$ and the positive rational numbers begins as

$$g(0) = 1/1 = 1, \quad g(1) = 1/2, \quad g(2) = 2/1 = 2,$$
$$g(3) = 1/3, \quad g(4) = 3/1 = 3,\ \ldots$$

($2/2 = 1$ was skipped since it was previously listed).

Now that we have a bijection g from $\mathbb{N}$ to the positive rational numbers, we can define a bijection $q : \mathbb{N} \to \mathbb{Q}$ as follows. We define $q(0) = 0$, $q(2n+1) = g(n)$, and $q(2n+2) = -g(n)$. Thus $\mathbb{Q}$ is a countably infinite set.

To summarize, we have shown the following theorem.

Theorem 1.1. *The following sets are countably infinite:* $\mathbb{N}$, $\mathbb{Z}$, *and* $\mathbb{Q}$.

Note that any infinite subset of a countably infinite set is also countably infinite. The elements of a countably infinite set may be listed as $a_1, a_2, a_3, \ldots$. Then the elements of an infinite subset may be listed as $a_{n_1}, a_{n_2}, a_{n_3}, \ldots$ for $n_1, n_2, n_3, \ldots$, an infinite subsequence of $1, 2, 3, \ldots$. Thus, for example, any infinite set of rational numbers is countably infinite.

If A and B are countably infinite, then $A \cup B$ is also countably infinite. This is seen as follows. For simplicity, we assume the sets are disjoint. Let $f : \mathbb{N} \to A$ and $g : \mathbb{N} \to B$ be bijective maps. Define $h : \mathbb{N} \to A \cup B$ by $h(2n) = f(n)$ and $h(2n+1) = g(n)$. It is easy to show that h is a bijective map.

There are numbers that are not rational numbers. These are called *irrational numbers*. For instance, $\sqrt{2}$ is irrational. To see this, suppose we have a rational number a/b, with the property that $(a/b)^2 = 2$. We may suppose that a/b is in lowest terms; that is, there is no common factor between a and b except 1. Then we get

$$a^2 = 2b^2,$$

showing us that the left-hand side is even. Thus a is even, and we can write $a = 2c$ for some integer c. We now get $4c^2 = 2b^2$, and cancelling the common factor of 2 on both sides of the equation yields

$$2c^2 = b^2.$$

This implies that the right-hand side is even, so that b is even. Thus we have both a and b are even, a contradiction.

$\sqrt{2}$ is an example of an *algebraic number*. A number α is said to be *algebraic* if α satisfies an equation of the form

$$\alpha^n + a_{n-1}\alpha^{n-1} + \cdots + a_1\alpha + a_0 = 0,$$

with a_i rational numbers. That is, the algebraic numbers are roots of polynomials with rational coefficients. Since $\sqrt{2}$ is a root of $x^2 - 2$, it is an algebraic number. One may show that the set of algebraic numbers is a countably infinite set, as is done in the exercises at the end of the chapter.

1.2. Uncountable Sets

From his musings on countable sets, Cantor went on to ask if $\mathbb{R}$, the set of real numbers, is countable. His first proof that the reals are uncountable, published in 1874, used nested intervals. His more famous proof, involving the *diagonal argument*, was published in 1891 and is given below.

Every real number x in the interval $(0, 1) = \{x \in \mathbb{R} : 0 < x < 1\}$ can be written as an infinite decimal:

$$x = 0.x_1x_2x_3\cdots .$$

Note that a decimal expansion ending in an infinite sequence of 0's $0.x_1x_2x_3\cdots x_{m-1}x_m000\cdots$ with $x_m \neq 0$ (called a *terminating expansion*) also has the expansion $0.x_1x_2x_3\cdots x_{m-1}y_m999\cdots$, where $y_m = x_m - 1$. If we agree never to allow an infinite sequence of 9's as the *tail* of the expansion, then the decimal expansion is unique. We can establish these assertions as follows. Take a real number $0 < x < 1$. Then $0 < 10x < 10$, so we may write

$$10x = x_1 + y_1,$$

where $0 \leq x_1 \leq 9$, $0 \leq y_1 < 1$, and x_1 is an integer. Then

$$\left| x - \frac{x_1}{10} \right| = \frac{y_1}{10} < \frac{1}{10}.$$

Iterate this procedure with y_1. Thus

$$10y_1 = x_2 + y_2,$$

where $0 \leq x_2 \leq 9$, $0 \leq y_2 < 1$, and x_2 is an integer. Thus

$$x = \frac{x_1}{10} + \frac{x_2}{10^2} + \frac{y_2}{10^2}.$$

Proceeding in this manner, we get

$$x - \frac{x_1}{10} - \frac{x_2}{10^2} - \cdots - \frac{x_n}{10^n} = \frac{y_n}{10^n},$$

where $0 \leq y_n < 1$. We see immediately that the decimal expansion converges to x. To establish uniqueness, let us suppose that

$$\sum_{n=1}^{\infty} \frac{x_n}{10^n} = \sum_{n=1}^{\infty} \frac{y_n}{10^n}$$

with $0 \leq x_n, y_n \leq 9$. Let m be the smallest number for which $x_m \neq y_m$. Without loss of generality, suppose that $x_m > y_m$. Then we have

$$\frac{x_m}{10^m} + \sum_{j=m+1}^{\infty} \frac{x_j}{10^j} = \frac{y_m}{10^m} + \sum_{j=m+1}^{\infty} \frac{y_j}{10^j}.$$

Thus

$$0 < \frac{x_m - y_m}{10^m} = \sum_{j=m+1}^{\infty} \frac{y_j - x_j}{10^j} \leq \frac{9}{10^{m+1}} \left(1 + \frac{1}{10} + \cdots \right) = \frac{1}{10^m}.$$

Hence $0 < x_m - y_m \leq 1$, which implies $x_m = y_m + 1$. Thus we must have $y_n = 9, x_n = 0$ for $n > m$. Thus uniqueness can fail only if one of our decimal expansions eventually ends in an infinite sequence of 9's.

We may now prove Cantor's theorem on the uncountability of $\mathbb{R}$.

Theorem 1.2. *The set $\mathbb{R}$ of real numbers is uncountable.*

Proof. Suppose that the real interval $(0,1)$ were countable. We may then list them:

$$\begin{aligned} r_1 &= 0.x_{11}x_{12}x_{13}\cdots \\ r_2 &= 0.x_{21}x_{22}x_{23}\cdots \\ &\vdots \end{aligned}$$

Now consider the number

$$r = 0.y_1y_2y_3\cdots,$$

where

$$y_n = \begin{cases} 1 & \text{if } x_{nn} \neq 1, \\ 2 & \text{if } x_{nn} = 1. \end{cases}$$

In this way, we avoid getting a 9 or 0 as a digit, thereby avoiding repeating 9's and ensuring $r \neq 0$. Then r is in $(0,1)$ but cannot appear in our listing above since it differs from each r_n in the nth digit. This is a contradiction, and hence the real interval $(0,1)$ is uncountable.

If a set A is uncountable and $A \subseteq B$, then B is also uncountable. To see this, suppose B were countable. Since A is infinite, B too is infinite, and hence countably infinite. Since A is an infinite subset of the countably infinite set B, it must be countably infinite, a contradiction. Thus, since $(0,1)$ is an uncountable subset of the real numbers, the set of all real numbers is uncountable. □

Suppose the set of irrational numbers were countable. Since $\mathbb{Q}$ is countably infinite, we would then have $\mathbb{R}$ as the union of two countable sets and hence countable, a contradiction. Thus there are uncountably many irrational numbers.

A real number that is not algebraic is called a *transcendental number*. Recall that the set of algebraic numbers is countably infinite. Suppose the set of transcendental numbers were countable. We would then have $\mathbb{R}$ as the union of two countable sets and hence countable, a contradiction. Thus there are uncountably many transcendental numbers.

In this sense, "most" real numbers are irrational, and in fact transcendental. Cantor showed the uncountability of the transcendental numbers in 1874. Before this, the only numbers known to be transcendental were numbers specifically constructed to be so (called Liouville numbers, named after Joseph Liouville (1809–1882)), and e, which was shown by Charles Hermite (1822–1901) to be transcendental just one year earlier. Thus Cantor proved that most real numbers are transcendental at a time when only a few examples were known! The transcendence of π was shown in 1882 by Lindemann. In his address to the ICM in 1900, Hilbert gave his list of twenty-three important unsolved problems in mathematics. In his seventh problem, he asked if a and b are algebraic numbers with $a \neq 0, 1$ and b irrational, does it follow that a^b is transcendental? The answer is yes, as was proved independently in 1934 by Alexander Gelfond (1906–1968) and Theodor Schneider (1911–1988). There are still many open questions regarding transcendental numbers. For example, we do not know if the numbers $\pi + e$ or πe are transcendental, although both are expected to be. It can be proved that at least one of them must be transcendental. This is an exercise in Chapter 7.

Instead of merely classifying sets as finite, countably infinite, and uncountable, we may refine this by saying that two sets A and B have the *same cardinality*, written $|A| = |B|$, if there is a bijective map between them. One may show that this is an equivalence relation. We say that A has *cardinality less than or equal to that of* B, written $|A| \leq |B|$, if there is an injective map from A to B. If there is an injective map from A to B but no bijective map between the sets is possible, we say A has *smaller cardinality* than B and write $|A| < |B|$. With Theorem 1.1 we showed that that $|\mathbb{N}| = |\mathbb{Z}| = |\mathbb{Q}|$. Since the inclusion map from $\mathbb{N}$ to $\mathbb{R}$ is injective, Theorem 1.2 shows that $|\mathbb{N}| < |\mathbb{R}|$.

Given a set A, consider its *power set* $P(A)$ defined as the set of all subsets of A. It is clear that the function $f : A \to P(A)$ defined by $f(a) = \{a\}$ is injective, and hence $|A| \leq |P(A)|$. Cantor proved the following theorem.

Theorem 1.3. *Let A be a set. There is no bijective map between A and $P(A)$, and hence $|A| < |P(A)|$.*

Proof. The proof is again by contradiction. Suppose there were a bijective map from A to $P(A)$. To each $a \in A$, we can then assign a unique set $T_a \in P(A)$. Consider

$$S = \{a \in A : a \notin T_a\}.$$

Clearly, S is a subset of A. Thus it must correspond to some T_w with $w \in A$. But this leads to a contradiction:

$$w \in S \implies w \notin T_w \implies w \notin S$$

and

$$w \notin S \implies w \in T_w \implies w \in S. \qquad \square$$

In this way, Cantor showed that there is an infinite ladder of infinite sets:

$$|\mathbb{N}| < |P(\mathbb{N})| < |P(P(\mathbb{N}))| < |P(P(P(\mathbb{N})))| < \cdots .$$

1.3. The Schröder–Bernstein Theorem

Instead of seeking a bijective correspondence between two sets A and B, it is sufficient to establish injective maps $f : A \to B$ and $g : B \to A$. In other words, if $|A| \leq |B|$ and $|B| \leq |A|$, then $|A| = |B|$. This is known as the Schröder–Bernstein[3] theorem after Ernst Schröder (1841–1902) and Felix Bernstein (1878–1956). Before proving this in general, we first prove a special case. If $B \subseteq A$, then the inclusion map from B to A is an injection. Thus $B \subseteq A$ implies $|B| \leq |A|$. The following lemma is the Schröder–Bernstein theorem in the special case where one set is a subset of the other.

Lemma 1.4. *Let A, B be sets such that $B \subseteq A$, and suppose that we have an injection $f : A \to B$. Then there is a bijection $g : A \to B$.*

[3]There is some controversy on the name of this theorem. It was first stated by Cantor without proof in 1887. It seems Richard Dedekind proved it in 1887 but didn't tell anyone about it. It was discovered in his notes in 1908. In 1895 Cantor published the first proof, but his proof uses the axiom of choice, which is discussed in the next chapter. Dedekind's unpublished proof did not use the axiom of choice. In 1896 Schröder published a proof sketch that was shown to be incorrect a few years later. In 1897, Bernstein proved the theorem—at age 19! Afterwards, Bernstein visited Dedekind, who apparently then independently proved the theorem yet again.

Proof. To prove this, we define sets $D_0, D_1, \ldots$ recursively as follows. $D_0 = A \backslash B$, $D_1 = f(D_0)$, $D_2 = f(D_1)$, and generally $D_{n+1} = f(D_n)$. Now define the map $g : A \to B$ by setting $g(x) = f(x)$ if x is in some D_n and $g(x) = x$ otherwise. If x is not in any D_n, then in particular it is not in D_0, and so x is in B so that $g(x) = x \in B$. We claim that g is a bijection. To see this, we have to show that g is injective and surjective. Suppose $g(x) = g(y)$. If both x and y are in some D_n, then we get $f(x) = f(y)$. Since f is injective, we deduce $x = y$. If both x and y are not in any D_n, then we have $x = g(x) = g(y) = y$, so again g is injective. Now consider the possibility that x is in some D_n and y is not. Then $g(x) = g(y)$ implies $f(x) = y$. Since x is in some D_n, it follows that y is in D_{n+1}, a contradiction. Thus g is injective. To see that g is surjective, let $b \in B$. If b is not in any D_n, then $g(b) = b$. If b is in some D_n, with $n \geq 1$, then $b \in f(D_{n-1})$ and so b is in the range of g. If $b \in D_0$, then $b \notin B$. □

Theorem 1.5 (Schröder–Bernstein theorem)**.** *If $f : A \to B$ and $g : B \to A$ are injective, then there is a bijection between A and B.*

Proof. The composition $g \circ f : A \to g(B)$ is also injective since

$$g(f(x)) = g(f(y)) \implies f(x) = f(y) \implies x = y.$$

$g(B)$ is a subset of A and so, by the previous lemma, there is a bijection $h : A \to g(B)$. Since g is injective, g^{-1} exists, and we have a map $g^{-1} : g(B) \to B$. Define $F : A \to B$ by $F(z) = g^{-1}(h(z))$. We show that F is both injective and surjective. If $F(z_1) = F(z_2)$, then $g^{-1}(h(z_1)) = g^{-1}(h(z_2))$, and applying g to both sides gives $h(z_1) = h(z_2)$. Since h is injective, we deduce $z_1 = z_2$. Let $b \in B$. There is an $a \in A$ such that $h(a) = g(b)$. Then $F(a) = g^{-1}(h(a)) = g^{-1}(g(b)) = b$. □

We have seen that $|\mathbb{N}| < |\mathbb{R}|$ and $|\mathbb{N}| < |P(\mathbb{N})|$. How do the cardinalities of $\mathbb{R}$ and $P(\mathbb{N})$ compare? We will use the Schröder–Bernstein theorem to prove that they are in fact the same.

Theorem 1.6. $|\mathbb{R}| = |P(\mathbb{N})|$.

Proof. Consider the function $f : P(\mathbb{N}) \to \mathbb{R}$ defined by $f(S) = 0.x_1x_2x_3\cdots$, where $x_i = 0$ if $i - 1 \in S$ and $x_i = 1$ if $i - 1 \notin S$.

Let $S, T \in P(\mathbb{N})$ with $f(S) = 0.x_1x_2x_3\cdots$ and $f(T) = 0.y_1y_2y_3\cdots$. Suppose $f(S) = f(T)$. Since we have avoided using the digit 9, the decimal expansions of $f(S)$ and $f(T)$ are unique. Thus $x_i = y_i$ for all $i \geq 1$, and in particular $x_i = 0$ if and only if $y_i = 0$. Thus $i - 1 \in S$ if and only if $i - 1 \in T$ for all $i \geq 1$, and so $S = T$. Thus f is injective, and hence $|P(\mathbb{N})| \leq |\mathbb{R}|$.

We now give an injective function mapping from $\mathbb{R}$ to $P(\mathbb{N})$. We may uniquely represent a real number x as $x = (-1)^n y + z$, where $n \in \{0, 1\}$, $y \in \mathbb{N}$, and $0 \leq z < 1$ (if $y = 0$, we agree to take $n = 0$). Furthermore, we may write $z = 0.d_1d_2d_3\cdots$ and agree to avoid the use of repeating 9's if the decimal expansion terminates. Let p_i be the ith prime number. We define $g : \mathbb{R} \to P(\mathbb{N})$ by $g(x) = \{2^n, 3^y\} \cup \{p_{i+2}^{d_i} : i \in \mathbb{N}\}$. For $x_1, x_2 \in \mathbb{R}$, we write $x_1 = (-1)^{n_1} y_1 + z_1$ and $x_2 = (-1)^{n_2} y_2 + z_2$ as described above. Suppose $g(x_1) = g(x_2)$. By the uniqueness of prime factorization, we have $n_1 = n_2$, $y_1 = y_2$, and all digits of z_1 and z_2 equal, and therefore $x_1 = x_2$. Thus g is injective, and hence $|\mathbb{R}| \leq |P(\mathbb{N})|$. By the Schröder–Bernstein theorem, $|\mathbb{R}| = |P(\mathbb{N})|$. □

Cantor's *continuum hypothesis* is the assertion that for any $\mathbb{N} \subseteq A \subseteq \mathbb{R}$, we either have $|A| = |\mathbb{N}|$ or $|A| = |\mathbb{R}|$. Cantor believed the hypothesis to be true and attempted, unsuccessfully, to prove it for many years. The problem was the first of Hilbert's aforementioned 1900 list of twenty-three unsolved problems. Through the work of Kurt Gödel in 1938 and Paul Cohen in 1963, the continuum hypothesis was shown to be independent of the usual axioms of set theory, which are discussed in the next chapter. The work of Gödel and Cohen is discussed in Section 4.3.

In 1884, the logician Gottlob Frege (1848–1925) defined an equivalence relation on the collection of sets by saying that two sets are equivalent if they have the same cardinality. The equivalence classes were to be thought of as *cardinal numbers*. Thus 0 represents the set of all sets with no elements, 1 represents the set of all sets with one element, and so forth.

However, it was quickly pointed out that the existence of certain large sets, such as the set of all sets with the same cardinality, may

lead to some fundamental difficulties. For example, consider the set of all sets that do not contain themselves. Does this set contain itself as an element? Either case yields a contradiction. This is the famous *Russell's paradox* which was brought to Frege's attention by Bertrand Russell (1872–1970). It may be rephrased as the *barber paradox*: a barber (who is male) shaves all men in his town who do not shave themselves. Does the barber shave himself? This self-reference is the fundamental obstacle. These paradoxes led mathematicians to re-examine the definition of a set and what the rules (or axioms) should be for constructing them. Thus it would be hoped that the creation of the set of all sets that do not contain themselves would not be allowed in such an axiomatic system. The most commonly used collection of axioms are discussed in the next chapter.

We end this section with a fun thought experiment, due to Hilbert. We describe an amazing hotel that we name *Hilbert's hotel* in honour of its creator. In this hotel, the number of rooms is countably infinite. There is a room for each positive natural number. Being a popular tourist destination, the hotel is full. On a rainy night, a poor wet soul stumbles into the lobby and pleads for a room. The attendant at the desk, feeling bad for the soaked patron, decides to make room even though the hotel is full. She gets on the hotel intercom and asks each guest to switch to the next room. If a guest is in room n, they are to move to room $n+1$. The hotel's guests are all very accommodating and happily oblige. Everyone still has a room, but now room 1 is free for the new guest!

Unfortunately, the desk attendant's work is not done for the night. A busload of new guests arrives and all need a separate room. However, this is a rather large bus, as it seats a countably infinite number of people! There is a seat for each positive natural number. Still, the crafty desk attendant is able to make room. On the hotel intercom, she asks all guests to move to the room that is twice their current room number. If a guest is in room n, they are to move to room $2n$. Doing this, all guests still have a room, but now all odd-numbered rooms are empty. The desk attendant then sends the person from bus seat n to room $2n-1$, which gives everyone a room and leaves the hotel full again.

Things can get even more difficult for our poor desk attendant. A fleet of busses arrives. There are a countably infinite number of busses in the fleet, one for each positive natural number, and each bus has a countably infinite number of seats! Our enterprising desk attendant again sends all current guests to the room twice their number, so that all even-numbered rooms are occupied. The occupant of the nth seat on the first bus is sent to room 3^n. The occupant of the nth seat on the second bus is sent to room 5^n. Similarly, the occupants of the third bus are sent to rooms that are powers of 7. The occupants of the next bus are sent to room that are powers of 11 (room $9 = 3^2$ is already occupied). In general, the nth occupant of bus m is sent to room number p_{m+1}^n, where p_m is the mth prime number (we skip powers of 2 since even-numbered rooms are occupied). The desk attendant is able to give everyone a room. Note that, unlike the previous strategies, this one leaves rooms unoccupied. For example, rooms 1 and 15 are left empty.

The example of Hilbert's hotel shows how counterintuitive infinite sets can be, especially if we are basing our expectations on the behaviour of finite sets.

Exercises

1.1. Show that the function $f : \mathbb{N} \to \mathbb{Z}$ given by

$$f(n) = (-1)^{n+1} \left\lfloor \frac{n+1}{2} \right\rfloor$$

is a bijection.

1.2. Give bijections $f : \mathbb{N} \to \mathbb{Z}$ and $P : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$, show that the function $h : \mathbb{Z} \times \mathbb{Z} \to \mathbb{N}$ defined by $h(x, y) = P(f^{-1}(x), f^{-1}(y))$ is bijective.

1.3. If A and B are countably infinite, show that $A \times B$ is countably infinite.

1.4. If A and B are countably infinite, show that $A \cup B$ is countably infinite.

1.5. If $A_1, A_2, \ldots,$ is an infinite sequence of disjoint finite sets, show that the union $\bigcup_{n=1}^{\infty} A_n$ is countably infinite.

1.6. If $A_1, A_2, \ldots$ is an infinite sequence of disjoint countably infinite sets, show that the union $\bigcup_{n=1}^{\infty} A_n$ is countably infinite. (In case you are aware of the *axiom of choice*, you may (and in fact will need to) use it here. The axiom will be discussed in the next chapter).

1.7. Construct an explicit polynomial bijection between $\mathbb{N} \times \mathbb{N} \times \mathbb{N}$ and $\mathbb{N}$.

1.8. Fix a natural number $b \geq 2$. Show that every positive real number x in $[0, 1]$ has a b-adic expansion of the form

$$x = \sum_{n=1}^{\infty} \frac{x_n}{b^n}$$

with each $0 \leq x_n \leq b - 1$ an integer.

1.9. Suppose

$$\sum_{n=1}^{\infty} \frac{x_n}{b^n} = \sum_{n=1}^{\infty} \frac{y_n}{b^n}$$

with each $0 \leq x_n \leq b - 1$ and $0 \leq y_n \leq b - 1$ integers. Show that either $x_n = y_n$ for all n, or there is an m such that one of the following two cases occurs:

- $x_m = y_m + 1$ and for $n \geq m + 1$, $y_n = b - 1$ and $x_n = 0$;
- $y_m = x_m + 1$ and for $n \geq m + 1$, $x_n = b - 1$ and $y_n = 0$.

1.10. Show that a number $x \in [0, 1]$ is rational if and only if its decimal expansion is eventually periodic. Deduce that irrational numbers have unique decimal expansions.

1.11. Show that the collection of polynomials with rational coefficients is a countably infinite set. Deduce that the set of algebraic numbers is countably infinite.

1.12. Show that the collection of infinite sequences made up of the elements 0 and 1 is uncountable. (*Hint*: Think about the proof that the set of real numbers between 0 and 1 is uncountable, and try something similar.)

1.13. Show that the number of functions mapping from $\mathbb{N}$ to $\mathbb{N}$ is uncountable. (*Hint*: Think about the proof that the number of

real numbers between 0 and 1 is uncountable, and try something similar.)

1.14. Define a relation on the collection of sets as follows. A is related to B if there is a bijection f mapping from A to B. Show that this is an equivalence relation. That is, show that the following hold.

- Any set A is related to itself.
- If A is related to B, then B is related to A.
- If A is related to B and B is related to C, then A is related to C.

1.15. Define $f : \mathbb{N}\times\mathbb{N} \to \mathbb{N}$ by $f(a, b) = 2^a 3^b$. Show that f is injective. Use the Schröder–Bernstein theorem to deduce that $\mathbb{N} \times \mathbb{N}$ is countably infinite.

1.16. Let A be the set of all finite subsets of $\mathbb{N}$. Find injective functions from $\mathbb{N}$ to A and from A to $\mathbb{N}$. (*Hint*: For the second function, try to use the uniqueness of prime factorization.) Use the Schröder–Bernstein theorem to deduce that A is countably infinite. Then prove that the number of infinite subsets of $\mathbb{N}$ is uncountable.

1.17. Let $\mathbb{R}^\times = \{x \in \mathbb{R} : x \neq 0\}$. Use the Schröder–Bernstein theorem to deduce that $|\mathbb{R}^\times| = |\mathbb{R}|$. Now try to explicitly define a bijection between the sets.

1.18. Let $A = \{x \in \mathbb{R} \mid 0 < x < 1\}$ and $B = \{x \in \mathbb{R} \mid 0 \leq x \leq 1\}$. Find injective functions $f : A \to B$ and $g : B \to A$. Use the Schröder–Bernstein theorem to deduce that $|A| = |B|$. Now try to explicitly define a bijection between the sets (this is a bit tricky).

1.19. Let $S = \{s_1, \ldots, s_n\}$ be a nonempty set of finitely many symbols. Show that the number of finite strings consisting of elements of S is countably infinite. What happens if S is countably infinite?

1.20. The two questions below refer to Hilbert's hotel, discussed at the end of this chapter.

(a) As in the final scenario, a fleet of a countably infinite number of busses arrives, each with a countably infinite number

of seats. Describe a way to assign rooms to everyone, including those currently in the hotel, so that no rooms are left empty.

(b) Multiple fleets of busses arrive at the hotel. There are a countably infinite number of fleets, one for each positive natural number. As before, each fleet contains a countably infinite number of busses, and each bus contains a countably infinite number of people. Find a way for the desk attendant to accommodate all guests.

Chapter 2

Axiomatic Set Theory

2.1. The Axioms

At the end of the previous chapter, we saw that the existence of certain sets leads to a mathematical contradiction. Thus we must be precise about the definition of a set, and we must have precise rules for constructing them. It is natural to think that if there is a property that applies to sets, then we should be able to construct a set of all sets with this property. Russell's paradox, mentioned earlier, concerns the set

$$R = \{s : s \notin s\}.$$

It is not difficult to see that $R \in R$ if and only if $R \notin R$, a contradiction.

This leads us to formalize the definition of a set carefully so that sets such as R (we hope) cannot be constructed. We do this by giving a list of *axioms*. These are statements that we will accept without proof, from which we hope to derive most, if not all, of mathematics. This chapter is intended to serve as an informal introduction to axiomatic set theory. Additional care must be taken when giving a rigorous description of formal set theory.

Our first axiom guarantees a nonempty universe. It is usually implied by the underlying formal logic. Since we have omitted discussing this here, we give it as an axiom.

Axiom 2.1 (Axiom of existence)**.**

$$\exists a(a = a).$$

We would like to know when two sets are equal. This leads to:

Axiom 2.2 (Axiom of extensionality)**.** If two sets share the same members, then they are equal. That is,

$$\forall x(x \in a \iff x \in b) \implies a = b.$$

A *formula* $\psi(x)$ in which x is a free variable is a logical statement whose truth depends on the value of x. In the next axiom, we try to give a rule for constructing a new set consisting of x that satisfy a formula $\psi(x)$. However, an axiom asserting the existence of a set b such that $x \in b \iff \psi(x)$ will lead us to Russell's paradox (just take $\psi(x)$ to be $x \notin x$). We avoid this by only allowing such a construction when all elements of b come from a previously constructed set.

Axiom 2.3 (Comprehension schema)**.** Given any set a and any formula $\psi(x)$ in which x is a free variable, there is a set b such that the members of b are precisely the members of a for which ψ holds. That is,

$$\exists b \forall x(x \in b \iff x \in a \ \wedge \ \psi(x)).$$

By extensionality, this set is unique. We will write

$$\{x \in a : \psi(x)\}$$

to designate the set b. This axiom is actually a collection of infinitely many axioms, one for each formula ψ. This is why it is referred to as a schema.

We may now deduce the existence of a set with no elements. By Axiom 2.1, a set a exists. By comprehension, we may form the set $\{x \in a : x \neq x\}$, which contains no elements. By extensionality, this set is unique. We designate it by $\varnothing$ and call it the *empty set*.

We may also now show that no universal set exists. Suppose u is a set containing all sets. By comprehension, we may form the set $\{x \in u : x \notin x\}$, which yields Russell's paradox, a contradiction. Thus, not every collection of sets is itself a set. It is still useful to be

able to discuss these collections of sets, and so we refer to them as *classes*. Informally, a class is "too big" to be a set.

Comprehension may be used to define the *intersection* of two sets: we define $a \cap b = \{x \in a : x \in b\}$. However, note that comprehension is not strong enough to allow us to define the union of two sets, as the union may not be a subset of a previously given set. We require some additional axioms to provide us with more methods for constructing sets.

Axiom 2.4 (Pairing axiom)**.** Given any two sets a and b, we can form a set that has exactly a and b as members. We write this as $\{a, b\}$.

We say a set a is a *subset* of b if $\forall x(x \in a \implies x \in b)$. When this happens, we write $a \subseteq b$.

Axiom 2.5 (Power set axiom)**.** Given a set a, there exists the *power set* of a, which is the collection of all of its subsets.

Axiom 2.6 (Union set axiom)**.** For any set a, there is a set, denoted $\bigcup a$, consisting of the elements of all elements of a. That is,

$$\forall x \, \forall y((x \in y \ \wedge \ y \in a) \implies x \in \bigcup a).$$

We may now define the *union* of two sets a and b. By the pairing axiom, $\{a, b\}$ is a set. We use the union set axiom to define $a \cup b = \bigcup\{a, b\}$.

We can construct the natural numbers from these axioms as follows. Since the empty set $\varnothing$ exists, we may use the pairing axiom to construct $\{\varnothing\}$, the set containing the empty set. We will call the empty set 0 and the set containing the empty set 1. By the pairing axiom, we can form $\{0, 1\}$. We will call this 2. We now have $2 = \{0, 1\}$ and, using the pairing axiom, we may form $\{2\}$. By the union set axiom, we have the set $2 \cup \{2\} = \{0, 1, 2\}$. We will call this 3. Proceeding recursively, we define the number $n + 1$ to be the set $n \cup \{n\}$. This allows us to construct the natural numbers from sets (in fact, from the empty set). We note that the usual $<$ ordering on the natural numbers is given by $\in$. That is, $m < n$ if and only if $m \in n$. Also, if m is an element of a natural number n, then m is

itself a smaller natural number, and hence all members of m are also members of n. That is, $m \in n$ implies $m \subseteq n$.

Although it is clear that there are infinitely many natural numbers, the previous axioms do not actually allow us to construct the set of all natural numbers. To do this, we require another axiom.

Axiom 2.7 (Axiom of infinity)**.** There is an infinite set. In particular, there is a set that contains $0 = \varnothing$, and for each natural number n in the set, $n + 1$ is in the set.

Using this axiom we may form ω, the set of all natural numbers.

These axioms are formalizing our intuitive notion that sets are built up out of smaller sets. Each set is composed of sets. Since sets are the building blocks, we do not want to have an infinite chain of sets within sets. We avoid this situation with the following axiom.

Axiom 2.8 (Axiom of regularity)**.** Any descending membership chain is finite.

Using this axiom, we can show that a set is never a member of itself. Indeed, if there is an a such that $a \in a$, then the infinite string of sets $a, a, \ldots$ violates the axiom of regularity. This axiom actually has no application in ordinary mathematics, in that ordinary mathematics may be done without using it. Removing it would present no real problems, although we keep it to prevent the possibility that $a \in a$, as this is not how we would typically imagine sets to work.

We say the formula $\psi(u, v)$ is a *function-like formula* if $\psi(u, v) \wedge \psi(u, w) \implies v = w$. Given a set a, we could like to form the set b containing all v for which there is a $u \in a$ such that $\psi(u, v)$ holds. It seems reasonable that such a set b would exist. We want to avoid constructing sets that are too big since they may lead to Russell's paradox. However, since $\psi(u, v)$ is a function-like formula, there are at most as many elements in b as there are in a. Note that we cannot use the comprehension schema to assert the existence of b since we are not requiring that the elements of b come from a previously constructed set. None of our previous axioms allow us to construct the set b, and so something new is required.

Axiom 2.9 (Replacement schema)**.**[1] Given a function-like formula $\psi(u, v)$ and a set a, there is a set b such that the members of b are exactly those sets v which correspond under ψ to some set u belonging to a. That is,

$$\exists b \forall v (v \in b \iff \exists u (u \in a \ \wedge \ \psi(u, v))).$$

We can define an *ordered pair* (a, b) to be the set $\{\{a\}, \{a, b\}\}$. One may show that $(a, b) = (c, d)$ if and only if $a = c$ and $b = d$ (see the Exercises). Our intuitive idea of a function can now be formalized. A function mapping from a set A to a set B is a set f of ordered pairs such that for each $a \in A$, there is a unique $b \in B$ such that $(a, b) \in f$. Writing $b = f(a)$, we have $f = \{(a, f(a)) : a \in A\}$.

The above list of axioms is the basis of Zermelo–Fraenkel set theory, denoted ZF. Ernst Zermelo (1871–1953) first gave a list of axioms for set theory in 1908. Abraham Fraenkel (1891–1965) proposed some modifications to these axioms, including the addition of the replacement schema. We will require one more axiom, originally proposed by Zermelo, called the axiom of choice.

Axiom 2.10 (Axiom of choice)**.** For any set a consisting of nonempty disjoint sets, there exists a set b such that for all $x \in a$, the intersection of b and x contains exactly one element.

In other words, given a set of nonempty sets, we can select a member from each set in this collection and put the selected elements together into a set. When we have a way to distinguish an element of each set in the collection, the axiom of choice is not needed to make this selection. However, the axiom allows us to make such a selection even when the elements of the sets are indistinguishable from each other.

Initially, the axiom of choice was controversial among mathematicians. Some were concerned that it implies the existence of a set in a nonconstructive sense. Others felt that some of its implications,

[1] This axiom schema can be used to deduce the pairing axiom and the comprehension schema. This is shown in the Exercises. We leave pairing and comprehension as separate axioms in this text in an effort to improve clarity.

such as the Banach–Tarski paradox,[2] were counterintuitive. However, criticism lessened when it was seen that the axiom had been subtly used in many previous results, including those by some of the axiom's harshest critics. Even though it is generally accepted today, it is still noted when it is included as an axiom of set theory. When the axiom of choice is included along with the axioms of ZF, they are denoted ZFC. In this book, we will be working in ZFC unless explicitly stated otherwise.

In 1938, Kurt Gödel (1906–1978) proved that the axiom of choice is consistent with ZF. That is, the negation of choice cannot be proved from the axioms of ZF. In 1963, Paul Cohen (1934–2007) proved that the negation of the axiom of choice is consistent with ZF. Thus the axiom of choice can neither be proved nor refuted from the axioms of ZF. When this happens, we say the axiom is *independent* of ZF. We discuss this in more detail near the end of Section 4.3.

2.2. Ordinal Numbers and Well Orderings

In the previous section, we have shown how the natural numbers can be constructed using the axioms of set theory. Once we have the natural numbers, the integers can be constructed. In particular, we define a relation on the set of ordered pairs of natural numbers so that (a, b) is related to (c, d) if $a + d = b + c$. One can show that this is an equivalence relation and define its equivalence classes to be the *integers*. The usual operations of arithmetic can then be extended to the integers, and their usual properties can be shown to hold. One can then define a relation on the set of ordered pairs of integers with nonzero second component so that (a, b) is related to (c, d) if $ad = bc$. Again, this is an equivalence relation, and one can define the *rational numbers* to be its equivalence classes. The usual operations of arithmetic can be extended to the rationals, and their usual properties can be shown to hold. It is a bit trickier to define the

[2]The Banach–Tarski paradox, named for Stefan Banach (1892–1945) and Alfred Tarski (1901–1983), is a theorem that says that a solid ball may be decomposed into finitely many pieces and then recomposed to form two identical copies of the original ball! In fact, it has been shown that only five pieces are needed. But don't go running for your saws and gold bars just yet; the fractal-like cuts required would be impossible to make in the physical universe.

set of real numbers. Typically, this is done with either Dedekind cuts or Cauchy sequences. Dedekind cuts are described in Section 7.1.

In this section, we would like to extend the natural numbers to somehow include *infinite numbers*. If we examine how we use the natural numbers, we see that we use them to both order things (for example, a house number in a postal address) and to count things. When we try to extend the natural numbers to include *infinite numbers*, these two notions no longer coincide. It is natural to first define *ordinal numbers*, and then to define *cardinal numbers* as certain ordinal numbers.

We have already seen that the notion of an ordered pair can be formulated in terms of sets. We define a *relation* R on a set S to be a set of ordered pairs of elements of S. We will say R is a *total ordering* of S if, writing $x < y$ to denote $(x, y) \in R$, the following hold:

(1) for any $x, y \in S$, we have $x = y$ or $x < y$ or $y < x$;

(2) for any $x \in S$, it is not the case that $x < x$ (R is irreflexive); and

(3) for any $x, y, z \in S$, $x < y$ and $y < z$ implies $x < z$ (R is transitive).

We say R *well-orders* S if R is a total ordering of S and if for every nonempty subset A of S, we have a *least element* in A with respect to R. For instance, the set of real numbers in $(0, 1)$ with respect to the usual definition of $<$ for real numbers is a total ordering but is not well-ordered. The set ω of all natural numbers is well-ordered under $\in$ (which, as noted above, corresponds to the usual $<$ ordering on the natural numbers). Each element of ω is also well-ordered under $\in$.

One may ask if every set admits a well-ordering. It turns out that this statement is equivalent to the axiom of choice. The axiom of choice has many equivalent forms. In fact, an entire book [**HR**] has been written on various statements equivalent to choice.

We say a set c is *transitive* if

$$\forall b(b \in c \implies b \subseteq c).$$

Note that this is equivalent to

$$\forall a\, \forall b((a \in b \ \wedge \ b \in c) \implies a \in c),$$

which is the usual transitive property on the $\in$ relation. ω is a transitive set, as is each element of ω.

Thus each natural number and ω itself are examples of transitive sets that are well-ordered by $\in$. We define a set α to be an *ordinal* if α is transitive and well-ordered by $\in$. This definition of ordinals is due to John von Neumann (1903–1957), who developed it at age 19.

Let α be an ordinal. The fact that we cannot have $\alpha \in \alpha$ follows from the axiom of regularity.[3] Suppose $X = \{\alpha : \alpha \text{ is an ordinal}\}$ were a set. In the Exercises, you will show that X is transitive and well-ordered by $\in$, and hence an ordinal itself. Thus $X \in X$, a contradiction. Therefore, the class of all ordinals is not a set.

Given sets S_1 and S_2 well-ordered by $<_1$ and $<_2$, respectively, we say the well-orderings are *order isomorphic* if there is a bijection between S_1 and S_2 that preserves the well-ordering. That is, if there is a bijection f from S_1 to S_2 with $x <_1 y \iff f(x) <_2 f(y)$. One may show that given any well-ordering on a set S, we can find an ordinal α that is order isomorphic to S. Thus, in a sense, the ordinals *are* the well-ordered sets (up to isomorphism).

So far we have constructed the natural numbers (finite ordinals) by setting 0 to be the empty set and setting $n + 1 = n \cup \{n\}$. We then set ω to be the set of all natural numbers and have seen that it too is an ordinal. We now set $\omega + 1 = \omega \cup \{\omega\}$. In general, given an ordinal α, we define the *successor ordinal* of α to be $\alpha + 1 = \alpha \cup \{\alpha\}$ (it is easy to check that $\alpha + 1$ is also an ordinal). Since $\alpha \in \alpha + 1$, we have $\alpha < \alpha + 1$.

Not every ordinal is a successor ordinal. Clearly 0 is not. The members of ω, the natural numbers, are all smaller ordinals. If ω were a successor ordinal, it too would be a natural number, which is not the case. Ordinals other than 0 that are not successor ordinals are called *limit ordinals*. Thus ω is an example of a limit ordinal.

We would like to define addition, multiplication, and exponentiation on ordinals. Given ordinals α and β, we can form the well-ordered sets $S = \{(0, x) : x \in \alpha\}$ and $T = \{(1, x) : x \in \beta\}$. We use S and T

[3]Note that even if the axiom of regularity were dropped, $\alpha \notin \alpha$ would still hold for ordinals due to $\in$ being a total ordering, and hence irreflexive. As previously mentioned, the axiom of regularity is not regularly used!

in place of α and β so that our sets are disjoint. We may now define a well-ordering on $S \cup T$ as follows: $(m, x) < (n, y)$ if $m < n$ or if $m = n$ and $x < y$. As noted above, this well-ordering will be order isomorphic to a unique ordinal; we take this ordinal to be $\alpha + \beta$. Informally, we form $\alpha + \beta$ by stacking β on top of α. We can check that this definition of addition does not contradict our earlier definition of $\alpha + 1 = \alpha \cup \{\alpha\}$, for here the ordering on $S \cup T$ yields

$$(0, 0) < (0, 1) < (0, 2) < \cdots < (1, 0).$$

We can identify $(0, x)$ with the element x in α, and $(1, 0)$ with α itself.

Ordinal addition is a bit peculiar. For example, $2 + \omega = \omega$, and yet $\omega + 2 \neq \omega$. To see the former, we note that the ordering on $S \cup T$ yields

$$(0, 0) < (0, 1) < (1, 0) < (1, 1) < (1, 2) < \cdots,$$

which is order isomorphic to ω. To see the latter, we note that the ordering on $S \cup T$ yields

$$(0, 0) < (0, 1) < (0, 2) < \cdots < (1, 0) < (1, 1),$$

which is clearly not order isomorphic to ω. For example, it has a greatest element, while ω does not.

Given ordinals α and β, we define a well-ordering on $\alpha \times \beta = \{(x, y) : x \in \alpha, y \in \beta\}$ as follows: $(x, y) < (t, u)$ if $y < u$ or if $y = u$ and $x < t$. Again, this well-ordering is isomorphic to a unique ordinal; we take this ordinal to be $\alpha\beta$. Ordinal multiplication is also a bit surprising at first. For example, we have $2\omega = \omega$ but $\omega 2 = \omega + \omega \neq \omega$. To see the former, we note that the ordering on $2 \times \omega$ yields

$$(0, 0) < (1, 0) < (0, 1) < (1, 1) < (0, 2) < (1, 2) < \cdots,$$

which is order isomorphic to ω. To see the latter, we note that the ordering on $\omega \times 2$ yields

$$(0, 0) < (1, 0) < (2, 0) < \cdots < (0, 1) < (1, 1) < (2, 1) < \cdots,$$

which is not order isomorphic to ω. For example, this ordering has two elements which do not have an immediate predecessor (namely $(0, 0)$ and $(0, 1)$), while ω has only one such element. However, when computing $\omega + \omega$, the ordering on $S \cup T$ yields

$$(0, 0) < (0, 1) < (0, 2) < \cdots < (1, 0) < (1, 1) < (1, 2) < \cdots,$$

which is clearly order isomorphic to the ordering for $\omega 2$ as given above. Thus $\omega 2 = \omega + \omega$. We note that $\omega 2$ is a limit ordinal.

It is a bit difficult to give a similar direct definition for ordinal exponentiation. Instead, given ordinals α and β, we define α^β recursively as follows:

(1) $\alpha^0 = 1$;

(2) $\alpha^{\beta+1} = \alpha^\beta \alpha$; and

(3) if β is a limit ordinal, then $\alpha^\beta = \bigcup\{\alpha^\gamma : \gamma \in \beta\}$.

As with addition and multiplication, ordinal exponentiation yields some seemingly odd results. For example, we have

$$2^\omega = \bigcup\{2^\gamma : \gamma \in \omega\} = \omega.$$

Although ordinal addition, multiplication, and exponentiation can yield some surprising results, they do satisfy many nice properties. For example, addition and multiplication are associative. That is, $(\alpha + \beta) + \gamma = \alpha + (\beta + \gamma)$ and $(\alpha\beta)\gamma = \alpha(\beta\gamma)$. The distributive law holds: $\alpha(\beta + \gamma) = \alpha\beta + \alpha\gamma$. Also, some familiar exponentiation laws hold, such as $\alpha^{\beta+\gamma} = \alpha^\beta \alpha^\gamma$ and $(\alpha^\beta)^\gamma = \alpha^{\beta\gamma}$.

One may show that the function $f(\alpha) = \omega^\alpha$ has ordinal fixed points, although the proof is beyond the scope of this book. We denote the smallest such fixed point by ϵ_0. It satisfies

$$\epsilon_0 = \bigcup\{\omega, \omega^\omega, \omega^{\omega^\omega}, \omega^{\omega^{\omega^\omega}}, \ldots\}.$$

In particular, we have $\omega^{\epsilon_0} = \epsilon_0$.

Recall that a set is countable if it is either finite or can be put in one-to-one correspondence with ω. Denote

$$\omega_1 = \{\alpha : \alpha \text{ is a countable ordinal}\}.$$

The replacement schema may be used to show that ω_1 is a set. It is straightforward to see that ω_1 is well-ordered by $\in$ and transitive, and hence an ordinal. If ω_1 were countable, we would have $\omega_1 \in \omega_1$, which cannot hold for ordinals (or any set, if we accept the axiom of regularity). Thus ω_1 is an uncountable ordinal. In fact, it may be shown to be the least uncountable ordinal. Although it seemed that

ϵ_0 is a very large ordinal number, it turns out that it is countable and hence $\epsilon_0 < \omega_1$.

Cantor showed that every ordinal may be written uniquely in the form

$$\omega^{\beta_1} c_1 + \omega^{\beta_2} c_2 + \cdots + \omega^{\beta_k} c_k,$$

where k and $c_1, \ldots, c_k$ are positive natural numbers, and $\beta_1 > \beta_2 > \cdots > \beta_k$ are ordinals. This representation is called the *Cantor normal form*.

We have created an increasing hierarchy of ordinal numbers, which essentially contain all well-ordered sets.

$$\begin{aligned} &0 < 1 < 2 < \cdots < \omega < \omega + 1 < \omega + 2 < \cdots < \omega + \omega \\ &= \omega 2 < \omega 3 < \cdots < \omega\omega = \omega^2 < \omega^3 < \cdots < \omega^\omega \\ &< \omega^{\omega^\omega} < \cdots < \epsilon_0 < \epsilon_0 + 1 < \cdots < \omega_1 < \omega_1 + 1 < \cdots . \end{aligned}$$

2.3. Cardinal Numbers and Cardinal Arithmetic

Another way of extending the natural numbers is with the notion of cardinality, which will lead us to *cardinal numbers*. We have already encountered this notion in Chapter 1. There, we said that two sets x and y have the *same cardinality* if there is a bijective correspondence between them, and we wrote $|x| = |y|$ when this happens. We wrote $|x| \leq |y|$ and said the cardinality of x is *less than or equal to* that of y if there is an injective map from x to y. If $|x| \leq |y|$ but there is no bijection from x to y, we wrote $|x| < |y|$. For example, for any $n \in \omega$, we have $|n| < |\omega|$. On the other hand, we have

$$|\omega| = |\omega + 1|.$$

This is because we may define a bijective correspondence between

$$\omega = \{0, 1, 2, \ldots\}$$

and

$$\omega + 1 = \{0, 1, 2, \ldots, \omega\}$$

by taking the mapping given by the maxim "the last shall be the first", and mapping ω in $\omega + 1$ to 0 and each $n \in \omega + 1$ with $n < \omega$ to $n + 1$.

In this way we may compare sizes of infinite sets. We would like to extend our definition of numbers to include infinite numbers that we will call the cardinalities of infinite sets. The ordinals themselves will not suffice, as we have just seen that distinct ordinals may have the same cardinality. As mentioned in Chapter 1, Frege attempted to define cardinal numbers with equivalence classes. He defined an equivalence relations on the collection of all sets by considering two sets equivalent if they have the same cardinality. However, this equivalence class cannot be a set. To see this, let us consider the equivalence class containing $1 = \{\varnothing\}$ (say) and suppose it were a set. This is the set that we would then hope to define as the cardinality of any one element set. Denoting this set by A, the union set axiom implies that $\bigcup A$ is a set. However, by the pairing axiom any set y is an element of the one element set $\{y\}$, and hence an element of $\bigcup A$. Thus $\bigcup A$ is a set consisting of all sets, which we have previously seen cannot exist. The equivalence class A is another example of a class that is too big to be a set. We would much prefer to have a definition of cardinal numbers in which they are sets, rather than classes.

By the axiom of choice, each set A may be well-ordered, and hence is order isomorphic to some ordinal α. We define the *cardinality* of A, denoted $|A|$, to be the least ordinal α order isomorphic to A, and refer to these ordinals as *cardinal numbers*. This definition of cardinal numbers is called the *von Neumann assignment*. We note that this definition of cardinality relies on the axiom of choice. Additional care must be taken to define cardinality without using the axiom of choice.

For example, the cardinality of any element n of ω is n, as we would expect. The cardinality of ω itself is the ordinal ω since any smaller ordinal is finite and hence not order isomorphic to ω. We denote this cardinality by the *cardinal* $\aleph_0$ (many sources instead use the notation ω_0). Thus we have seen that $|\omega + 1| = |\omega| = \aleph_0$ (and, in fact, $\aleph_0 = \omega$). The argument we used to show this may be generalized: for any infinite cardinal number κ, its ordinal successor $\kappa + 1$ has cardinality κ. In Chapter 1 we saw that $|\mathbb{N}| = |\mathbb{Z}| = |\mathbb{Q}| = \aleph_0$.

We note that when writing $\kappa < \lambda$ for cardinals κ and λ, there is no confusion on whether we are referring to the ordinal or cardinal inequality, as their meaning is actually equivalent. Since κ and λ are

ordinals, $\kappa < \lambda$ means that $\kappa \in \lambda$. Since $|\kappa| = \kappa$ and $|\lambda| = \lambda$, $\kappa < \lambda$ means that $|\kappa| < |\lambda|$, which is to say that there is an injective map from κ to λ, but no bijective map. One may show that these two notions coincide (this is an Exercise).

Given a cardinal number κ, we define the *successor cardinal* κ^+ to be the least cardinal number greater than κ. We define the *cardinals* $\aleph_\alpha$ by recursion on the ordinal α:

(1) $\aleph_0 = \omega$;

(2) $\aleph_{\alpha+1} = (\aleph_\alpha)^+$;

(3) if γ is a limit ordinal, then $\aleph_\gamma = \bigcup\{\aleph_\alpha : \alpha \in \gamma\}$.

One may show that the cardinals $\aleph_\alpha$ are each cardinal numbers and, in fact, represent all possible cardinal numbers. Since we have previously seen that

$$\omega_1 = \{\alpha : \alpha \text{ is a countable ordinal}\}$$

is the least uncountable ordinal, it follows that $\aleph_1 = \omega_1$. This notation for cardinals is due to Cantor.

If the notion of cardinality is to be meaningful, we should be able to compare the sizes of any two sets. That is, given any two cardinals κ and λ, we must have

$$\kappa < \lambda, \quad \lambda < \kappa, \quad \text{or} \quad \kappa = \lambda.$$

This is called the *trichotomy law for cardinals*. Since κ and λ are ordinals, this is equivalent to the statement that

$$\kappa \in \lambda, \quad \lambda \in \kappa, \quad \text{or} \quad \kappa = \lambda,$$

which is true given our construction of ordinals. In fact, this gives us an easy proof of the Schröder–Bernstein theorem. Suppose A and B are infinite sets. By the axiom of choice, they can be well-ordered and hence are order isomorphic to ordinals α and β, respectively. If $|A| \leq |B|$ and $|B| \leq |A|$ but $|A| \neq |B|$, it would follow that $\alpha \in \beta$ and $\beta \in \alpha$, which cannot hold. The advantage of the proof of the Schröder–Bernstein theorem given in Exercise 1 is that it does not require the use of the axiom of choice.

We would now like to define addition, multiplication, and exponentiation on cardinals. Although each cardinal is an ordinal, the

operations of cardinal arithmetic will be distinct from those of ordinal arithmetic. If A and B are sets, then their *cartesian product* is

$$A \times B = \{(a, b) : a \in A \wedge b \in B\}.$$

Given cardinals κ and λ, we define their sum and product as

$$\kappa + \lambda = |\kappa \times \{0\} \cup \lambda \times \{1\}|,$$

$$\kappa \cdot \lambda = |\kappa \times \lambda|.$$

The cartesian product is used in the sum in order to ensure the two sets in the union are disjoint. Basically, the sum of two cardinals is the cardinality of the union of two disjoint copies of the cardinals, and the product of the cardinals is the cardinality of their cartesian product. One can easily check that these definitions correspond to the usual definitions of addition and multiplication on the finite cardinals. If at least one of κ or λ is infinite and the other is nonzero, one may show that

$$\kappa + \lambda = \kappa \cdot \lambda = \max\{\kappa, \lambda\}.$$

Thus cardinal addition and multiplication are rather trivial operations, in a sense.

Things are not as trivial for cardinal exponentiation. We define

$$\kappa^\lambda = |\{f : f \text{ is a function from } \lambda \text{ to } \kappa\}|.$$

It is easy to check that this corresponds to the usual definition of exponentiation when κ and λ are finite cardinals. We note that for any cardinal κ, if $|A| = \kappa$, then $|P(A)| = 2^\kappa$. To show this, we must construct a bijective correspondence between the power set of A and the set of all functions mapping from A to $2 = \{0, 1\}$. This may be done by corresponding a subset B of A with the function $f_B : A \to \{0, 1\}$ that sends elements of B to 1 and all other elements to 0.

We may rewrite some results from Chapter 1 using cardinal numbers. Theorem 1.3, Cantor's power set theorem, implies $\kappa < 2^\kappa$ for any cardinal κ. Theorem 1.6 showed that $|\mathbb{R}| = |P(\mathbb{N})|$. This may now be written as $|\mathbb{R}| = 2^{\aleph_0}$. The continuum hypothesis states that $2^{\aleph_0} = \aleph_1$. As mentioned in Chapter 1, this statement is independent

of the axioms of ZFC set theory. In this sense, cardinal exponentiation is highly nontrivial, as the ZFC axioms cannot even allow us to determine which $\aleph_\alpha$ is equal to $2^{\aleph_0}$![4] The *generalized continuum hypothesis* states that for any ordinal α, we have $\aleph_{\alpha+1} = 2^{\aleph_\alpha}$. It too is independent of the ZFC axioms by the works of Gödel and Cohen. However, it is not independent of the ZF axioms: in 1947, Wacław Sierpiński (1882–1969) proved that the generalized continuum hypothesis with ZF implies the axiom of choice.

Further Reading

Jech's *Set Theory* [**Je**] and Kunen's *Set Theory* [**Ku**] give comprehensive treatments of the topic. However, they may be a bit difficult to use as an introduction to set theory. Goldrei's *Classic Set Theory for Guided Independent Study* [**Go**] is a good introduction to the material. Paul Halmos' *Naive Set Theory* [**Ha**] is another popular introduction.

Exercises

2.1. Recall that an ordered pair (a, b) can be defined as the set $\{\{a\}, \{a, b\}\}$. Show that $(a, b) = (c, d)$ if and only if $a = c$ and $b = d$.

2.2. Define the ordered triple (a, b, c) to be the ordered pair $((a, b), c)$, where the ordered pair is defined as usual. Show that

$$(a_1, b_1, c_1) = (a_2, b_2, c_2)$$

if and only if $a_1 = a_2$, $b_1 = b_2$, and $c_1 = c_2$.

2.3. Show that the replacement schema implies the comprehension schema.

2.4. In this question, we show how the pairing axiom follows from the replacement schema. Let sets a and b be given.

[4] We can rule out some possibilities. For example, it can be shown that $2^{\aleph_0}$ cannot be equal to $\aleph_\omega$.

(a) We originally used the pairing axiom to construct the set $\{\varnothing, \{\varnothing\}\}$ (which was then defined to be the number 2). Instead, use the power set axiom to construct this set.

(b) Let $\psi(u, v)$ be the formula

$$(u = \varnothing \ \wedge \ v = a) \quad \text{or} \quad (u \neq \varnothing \ \wedge \ v = b).$$

Show that this is a function-like formula.

(c) Use the replacement schema on the set $\{\varnothing, \{\varnothing\}\}$ and the function-like formula $\psi(u, v)$ to show the existence of the set with elements a and b.

2.5. (a) Define a relation on the set of ordered pairs of natural numbers as follows: (a, b) is related to (c, d) if $a+d = b+c$. Show that this is an equivalence relation. You may assume that addition on the natural numbers is commutative, and that if $x + z = y + z$, then $x = y$.

(b) Let S be the set of ordered pairs of integers with a nonzero second component. Define a relation on S as follows: (a, b) is related to (c, d) if $ad = bc$. Show that this is an equivalence relation. You may assume that multiplication on the integers is commutative, that if $xz = yz$ and $z \neq 0$, then $x = y$, and that $0x = 0$ for any x.

2.6. Suppose $X = \{\alpha : \alpha \text{ is an ordinal}\}$ were a set. Show that it would follow that X is transitive and well-ordered by $\in$.

2.7. Suppose α is an ordinal. Show that $\alpha \cup \{\alpha\}$ is also an ordinal.

2.8. Let α and β be ordinals, and let $S = \{(0, x) : x \in \alpha\}$ and $T = \{(1, x) : x \in \beta\}$. Define an ordering on $S \cup T$ as follows: $(m, x) < (n, y)$ if $m < n$, or if $m = n$ and $x < y$. Show that this is a well-ordering of $S \cup T$.

2.9. Let α and β be ordinals. We define an ordering on $\alpha \times \beta = \{(x, y) : x \in \alpha, y \in \beta\}$ as follows: $(x, y) < (t, u)$ if $y < u$, or if $y = u$ and $x < t$. Show that this is a well-ordering of $S \times T$.

2.10. Given two ordinals α and β, we define $\alpha + \beta$ and $\alpha\beta$ to be the ordinals that are order isomorphic to the well-ordered sets defined in the previous two questions. With ω denoting the first infinite ordinal, show that $\omega + 2 \neq 2 + \omega$ and $2\omega \neq \omega 2$.

2.11. Let ω_1 be the set of all countable ordinal numbers. Show that ω_1 is an ordinal.

2.12. Let κ and λ be cardinals (and hence ordinals). Show that $\kappa \in \lambda$ if and only if there is an injective function from κ to λ, but there does not exist a bijective function between κ and λ. This shows that the ordinal inequality $\kappa < \lambda$ holds if and only if the cardinal inequality $\kappa < \lambda$ holds.

2.13. Let A be a set. Given a subset B of A, define $f_B : A \to \{0, 1\}$ by

$$f_B(x) = \begin{cases} 1 & \text{if } x \in B, \\ 0 & \text{if } x \notin B. \end{cases}$$

Let C be the set of all functions mapping from B to $\{0, 1\}$, and define $\Phi : P(A) \to C$ by $\Phi(B) = f_B$. Show that Φ is bijective. This shows that if $|A| = \kappa$, then $|P(A)| = 2^\kappa$.

Chapter 3

Elementary Number Theory

3.1. Divisibility

The word *algorithm* comes from the name of the 9th century Persian mathematician Muḥammad ibn Mūsā al-Khwārizmī who wrote a mathematical work around 820 CE entitled *Al-kitāb al-mukhtaṣar fī ḥisāb al-jabr wa'l-muqābala*.[1] However, algorithms existed long before this work. We will explore algorithms more carefully in Chapter 4. For now, we may informally define an algorithm as a set of clear instructions for solving a certain type of problem.

The most familiar of algorithms is Euclid's algorithm for computing the greatest common divisor of two natural numbers. Written around 300 BCE, this algorithm, which will be given below, is based on the *division algorithm*.[2] Given two integers a and b with $b \neq 0$, there exist a quotient $q \in \mathbb{Z}$ and a remainder $r \in \mathbb{Z}$ with

$$a = qb + r \quad \text{and} \quad 0 \leq r < b.$$

[1] The word *algebra* is also derived from 'al-jabr' appearing in the title of this book.

[2] One could make the case that the division algorithm (at least as we state it here) is poorly named, since it merely claims the existence of certain integers and does not tell how to find them. The division algorithm is not an algorithm! Of course, it is not too difficult to describe an actual algorithm that will compute the quotient and remainder, such as long division.

If $r = 0$, we say b *divides* a and call b a *divisor* of a, and a a *multiple* of b. If b divides a, we write $b \mid a$.

The concept of division gives a relation on the integers. There are several properties of this relation that we can write down immediately. For example, if $c \mid a$ and $c \mid b$, then $c \mid (a + b)$. Also, if $b \mid a$ and k is an integer, then $b \mid ka$. Finally, if $b \mid a$ and $a \mid b$, then $a = \pm b$. These properties can be easily verified by the reader.

Given any positive integer a, there are only finitely many divisors of a. Given any two natural numbers a and b, not both 0, we can consider the divisors of a and the divisors of b. We can take the intersection of these two sets and ask for the greatest element. Since at least one of a or b is nonzero, this intersection is finite. Since 1 is a divisor of both a and b, the intersection is nonempty, and thus there is a greatest element, which we call the *greatest common divisor*, or *gcd*, of a and b. We denote this by $\gcd(a, b)$.

Euclid observed that if we have two natural numbers a and b with $0 < b < a$, and we want to compute the greatest common divisor of a and b, we can reduce the size of our numbers with the following observation. Let $d = \gcd(a, b)$. By the division algorithm, we have $a = qb + r$, and we see that $d = \gcd(b, r)$. Indeed, $d \mid \gcd(b, r)$ but $\gcd(b, r) \mid d$ also. The key point in this observation is that $b < a$ and $0 \leq r < b$, and so the size of our pair of numbers has been reduced. We can iterate this procedure until we get a remainder of zero. As an example, let us compute $\gcd(8255, 3556)$.[3] We have

$$\begin{aligned} 8255 &= 2 \cdot 3556 + 1143, \\ 3556 &= 3 \cdot 1143 + 127, \\ 1143 &= 9 \cdot 127 + 0. \end{aligned}$$

[3]Of course, the greatest common divisor may be found via factorization. Since $8255 = 5 \cdot 13 \cdot 127$ and $3556 = 2^2 \cdot 7 \cdot 127$, we have $\gcd(8255.3556) = 127$. However, finding the prime factorization of a number generally takes a long time, especially for large numbers. The Euclidean algorithm gives a much faster method of computing the greatest common divisor. This fact is very important in many computing applications, such as RSA cryptography.

Thus

$$\begin{aligned}\gcd(8255, 3556) &= \gcd(3556, 1143)\\ &= \gcd(1143, 127)\\ &= \gcd(127, 0)\\ &= 127.\end{aligned}$$

This procedure is called the *Euclidean algorithm*. Since any decreasing sequence of natural numbers must eventually terminate, the Euclidean algorithm will return the greatest common divisor after finitely many steps.

There is actually more that can be deduced from this procedure. From the equation $r = a - qb$, we see that r can be written as a linear combination of a and b. From $b = q_1 r + r_1$, we see that r_1 is a linear combination of r and b. Thus r_1 is a linear combination of a and b. Continuing in this way, we deduce the following theorem:[4]

Theorem 3.1. *Let a and b be positive natural numbers. Then there are integers x_0 and y_0 with*

$$\gcd(a, b) = ax_0 + by_0.$$

This idea of "working backwards" can be stated clearly in matrix form. We have

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} q & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} b \\ r \end{pmatrix} = \begin{pmatrix} q & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} q_1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} r \\ r_1 \end{pmatrix}.$$

Iterating this procedure, we see that

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} q & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} q_1 & 1 \\ 1 & 0 \end{pmatrix} \cdots \begin{pmatrix} q_k & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} d \\ 0 \end{pmatrix},$$

where $d = \gcd(a, b)$. Each of the two-by-two matrices appearing on the right-hand side are invertible, and thus we can solve for d. Multiplying by these inverses, we have

$$\begin{pmatrix} d \\ 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -q_k \end{pmatrix} \cdots \begin{pmatrix} 0 & 1 \\ 1 & -q_1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & -q \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}.$$

[4] According to Weil [**We**], this theorem is due to Aryabhata. On pages 6 to 7, Weil writes "Indeterminate equations of the first degree, to be solved in integers, must have occurred quite early in various cultures... if we leave China aside, the first explicit description of the general solution occurs in the mathematical portion of the Sanskrit astronomical work Āryabhaṭīya, of the fifth–sixth century A.D." See also [**Du**].

For example, using our numbers in the example above, we have

$$\begin{aligned}\begin{pmatrix}127\\0\end{pmatrix} &= \begin{pmatrix}0&1\\1&-2\end{pmatrix}\begin{pmatrix}0&1\\1&-3\end{pmatrix}\begin{pmatrix}0&1\\1&-6\end{pmatrix}\begin{pmatrix}8255\\3556\end{pmatrix}\\ &= \begin{pmatrix}-3&7\\28&-65\end{pmatrix}\begin{pmatrix}8255\\3556\end{pmatrix},\end{aligned}$$

and so $7 \cdot 3556 - 3 \cdot 8255 = 127$.

We can formalize these observations in the following theorem.

Theorem 3.2. *Given two positive natural numbers a and b, let $d = \gcd(a, b)$. The equation*

$$ax + by = c$$

has a solution in integers if and only if $d \mid c$.

Proof. The necessity is clear since d divides the left-hand side and so must divide the right-hand side. To establish sufficiency, we can use Theorem 3.1 to obtain integers x_0 and y_0 satisfying $ax_0 + by_0 = d$ and then multiply both sides by the integer c/d. □

When two numbers have a greatest common divisor of 1, we say they are *relatively prime*, or *coprime*. We note that if a and b have a greatest common divisor of d, then a/d and b/d are relatively prime. The following lemma will be useful.

Lemma 3.3. *If $a \mid bc$ and $\gcd(a, b) = 1$, then $a \mid c$.*

Proof. Since $\gcd(a, b) = 1$, by Theorem 3.2 there are integers x, y such that $ax + by = 1$. Multiplying this by c, we get $acx + bcy = c$. Since $a \mid ac$ and $a \mid bc$, we deduce $a \mid (acx + bcy)$, and thus $a \mid c$. □

With this lemma, we can now prove an important result.

Theorem 3.4. *Given two positive natural numbers a and b, let $d = \gcd(a, b)$. Suppose that $d \mid c$. Let x_0, y_0 satisfy $ax_0 + by_0 = c$. Then all the solutions of*

$$ax + by = c$$

are of the form $x = x_0 + (b/d)t$ and $y = y_0 - (a/d)t$, with t ranging over all integers.

Proof. It is clear that any x and y of this form gives a solution of the equation. We note that x_0 and y_0 satisfying $ax_0+by_0=c$ must exist by Theorem 3.2. If $ax+by=c$, we can subtract and deduce $a(x-x_0)+b(y-y_0)=0$. Dividing by d, we get $(a/d)(x-x_0)=(b/d)(y_0-y)$. Since a/d divides the left-hand side and $\gcd(a/d, b/d)=1$, we deduce by Lemma 3.3 that $(a/d) \mid (y_0-y)$. Thus there is an integer t such that $(a/d)t=y_0-y$. Putting this back into our equation yields

$$(a/d)(x-x_0)=(b/d)(y_0-y)=(b/d)(a/d)t,$$

which implies $x-x_0=(b/d)t$. □

This theorem allows us to determine in an effective manner all the integer solutions of an equation $ax+by=c$.

A natural number p is called a *prime number* if its positive divisors are only 1 and itself. We exclude 1 from the prime numbers (we will see the reason for this exclusion soon). Thus $2, 3, 5, 7, 11, 13, \ldots$ are the first few prime numbers. We begin by proving a basic property of prime numbers.

Lemma 3.5. *If p is a prime number with $p \mid ab$, then $p \mid a$ or $p \mid b$.*

Proof. Suppose p does not divide a. Then p and a are relatively prime, so by Theorem 3.2 we can find integers x and y such that $px+ay=1$. Multiplying this equation by b yields $pbx+aby=b$. Since p divides the left-hand side, we have $p \mid b$. □

With this result in hand, we can proceed to prove the *fundamental theorem of arithmetic.*[5]

Theorem 3.6 (Fundamental theorem of arithmetic)**.** *Every natural number greater than* 1 *can be written as a product of prime numbers, and this factorization is unique.*

Proof. We first show that every natural number greater than 1 is divisible by a prime number. We proceed by induction. If n is not prime, then we can write $n=ab$ with $1<a<n$ and $1<b<n$. By induction, each of a and b are divisible by a prime number and hence

[5]Surprisingly, the fundamental theorem of arithmetic does not appear in Euclid's *Elements*. It was first written down in 1801 by Carl Friedrich Gauss (1777–1855) in his *Disquisitiones Arithmeticae*. See p. 5 of [**We**].

n has a prime divisor. Factorization of every natural number greater than 1 into a product of prime numbers again follows by an induction argument. Uniqueness follows from Lemma 3.5 since if

$$p_1^{a_1} \cdots p_k^{a_k} = q_1^{b_1} \cdots q_t^{b_t}$$

with the p_i and q_j prime, it follows that p_1 divides the left-hand side and hence must divide the right-hand side. Suppose without loss of generality that $k \leq t$. By Lemma 3.5, $p_1 \mid q_j$ for some j. Since q_j is prime, it has only two divisors, namely 1 and q_j. Thus, $p_1 = q_j$ for some j. Proceeding in this way, we pair up all the primes on the left-hand side with those of the right-hand side, and uniqueness is established. □

This theorem is the reason we do not consider 1 to be a prime. If 1 were allowed to be a prime, then prime factorization would no longer be unique.

An important consequence of unique factorization is the following.

Lemma 3.7. *If $ab = c^2$ and $\gcd(a, b) = 1$, then a and b are perfect squares.*

Proof. We factor a and b as products of prime powers. Since the factorization is unique, we obtain the unique factorization of c^2 also. But this means that each prime is raised to an even power in the factorization. Each prime power divides either a or b but not both, since a and b are relatively prime. It follows that a and b are perfect squares. □

We write $a \equiv b \pmod{m}$ to mean that $m \mid (a - b)$, and we say a is *congruent* to b modulo m. Congruence modulo m is an equivalence relation, and its equivalence classes are called *residue classes*. Every number is congruent to its remainder when divided by m, and so every number is congruent to one of $0, 1, \ldots, m-1$. Furthermore, no two of these numbers can be congruent modulo m, since their difference will be a nonzero number less than m and greater than $-m$, and hence not divisible by m.

We note that if a number x is even, then we can write $x = 2k$, and so $x^2 = 4k^2 \equiv 0 \pmod 4$. If x is odd, then we can write $x = 2k + 1$, and so $x^2 = 4k^2 + 4k + 1 \equiv 1 \pmod 4$. Thus squares are congruent to 0 or 1 modulo 4, according to whether they are even or odd.

As an interesting application of some of the results we have seen in this section, we can characterize integer solutions to the equation $x^2 + y^2 = z^2$, called the *Pythagorean triples*.

Theorem 3.8. *Suppose x, y, and z are integers such that*

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

Then there exist integers a, b, c such that

$$\{x, y\} = \{c(a^2 - b^2), 2cab\} \quad \textit{and} \quad z = c(a^2 + b^2).$$

Conversely, any x, y, and z with a, b, c defined as above satisfy $x^2 + y^2 = z^2$.

Proof. Let us first consider the *primitive* Pythagorean triples, those with $\gcd(x, y, z) = 1$. For any such triples, x and y cannot both be even, for then z would also be even. If x and y are both odd, then $x^2 \equiv 1 \pmod 4$ and $y^2 \equiv 1 \pmod 4$, and hence $z^2 = x^2 + y^2 \equiv 2 \pmod 4$, a contradiction. Thus, without loss of generality, we may suppose x is odd and y is even, which implies z is odd. Put $y = 2y_1$. Then

$$y_1^2 = \left(\frac{z - x}{2}\right)\left(\frac{z + x}{2}\right).$$

The two integer factors on the right are relatively prime, for suppose otherwise. If p is a common prime factor, then p divides both their sum and difference. That is, $p \mid z$ and $p \mid x$, which would mean p divides $y^2 = z^2 - x^2$, and hence $p \mid y$, contrary to the primitive assumption. Thus by Lemma 3.7 each factor of y_1^2 must be a perfect square. Hence

$$\frac{z - x}{2} = b^2 \quad \text{and} \quad \frac{z + x}{2} = a^2.$$

Taking the sum, difference, and product yields $z = a^2 + b^2$, $x = a^2 - b^2$, and $y = 2ab$. Conversely, any such triple satisfies $x^2 + y^2 = z^2$. This completely determines the primitive Pythagorean triples. Now, if $c = \gcd(x, y, z)$, then we may write $x = cx_1$, $y = cy_1$, and $z = cz_1$ so

that x_1, y_1, and z_1 form a primitive Pythagorean triple, and these are characterized by the above discussion. This completes the proof. □

We end this section with a few results on modular arithmetic. The first, about the existence of multiplicative inverses, follows quickly from Theorem 3.2.

Lemma 3.9. *Let a and n be numbers with $n \geq 2$ and $\gcd(a, n) = 1$. Then there exists a number b with $ab \equiv 1 \pmod{n}$. Furthermore, if we also have $ac \equiv 1 \pmod{n}$, then $b \equiv c \pmod{n}$.*

Proof. By Theorem 3.2, there exist integers x and y with $ax+ny = 1$. Reducing modulo n yields the result. To show uniqueness of the inverse modulo n, suppose $ab \equiv 1 \equiv ac \pmod{n}$. Then $n \mid a(b-c)$. Since $\gcd(a, n) = 1$, Lemma 3.3 implies $n \mid (b-c)$, and so $b \equiv c \pmod{n}$. □

We now give a result known as *Fermat's little theorem*.[6]

Theorem 3.10 (Fermat's little theorem)**.** *Let p be a prime, and let a be an integer. If p does not divide a, then $a^{p-1} \equiv 1 \pmod{p}$.*

Proof. Suppose p does not divide a. Then for $0 < k < \ell < p$, we claim $ka \not\equiv \ell a \pmod{p}$, for suppose otherwise. Then $p \mid a(\ell - k)$. Lemma 3.3 then implies $p \mid (\ell - k)$, which is impossible since $0 < \ell - k < p$. It follows from the above claim that $1a, 2a, \ldots, (p-1)a$ form a complete set of nonzero mutually incongruent residues modulo p. Since $1, 2, \ldots, (p-1)$ also form such a set, we must have

$$1a \cdot 2a \cdots (p-1)a \equiv 1 \cdot 2 \cdots (p-1) \pmod{p},$$

and hence $(p-1)!a^{p-1} \equiv (p-1)! \pmod{p}$. Since $\gcd((p-1)!, p) = 1$, Lemma 3.9 implies $(p-1)!$ is invertible modulo p. Multiplying by this inverse yields the result. □

We now give *Wilson's theorem*, which gives us a way to characterize the prime numbers that will be useful to us later on.

[6] A generalization of Fermat's little theorem to composite moduli is due to Euler. See Exercise 3.6 at the end of this chapter.

Theorem 3.11 (Wilson's theorem)**.** *A natural number $n > 1$ is prime if and only if*

$$(n-1)! \equiv -1 \pmod{n}.$$

Proof. Suppose n is composite. Then there is a prime q dividing n with $1 < q < n$. Hence $q \mid (n-1)!$ and so $(n-1)! \equiv 0 \pmod{q}$. Then $(n-1)! \equiv -1 \pmod{n}$ with $q \mid n$ would imply $(n-1)! \equiv -1 \pmod{q}$, a contradiction. Conversely, suppose $n = p$ is prime. The case $p = 2$ is easily verified, and so we may assume $p \geq 3$. By Lemma 3.9, for each a with $1 \leq a < p$, there exists a unique b with $1 \leq b < p$ such that $ab \equiv 1 \pmod{p}$. Are there a that serve as their own inverse? Suppose $a^2 \equiv 1 \pmod{p}$. Then $p \mid (a-1)(a+1)$, and so Lemma 3.5 implies $p \mid (a-1)$ or $p \mid (a+1)$. That is, $a \equiv 1 \pmod{p}$ or $a \equiv -1 \equiv p-1 \pmod{p}$. Thus each of the numbers $1, 2, \ldots, p-1$ can be paired up with their distinct inverse, except for 1 and $p-1$. Hence $(p-1)! \equiv (1)(p-1) \equiv -1 \pmod{p}$. □

To end this section, we discuss the *Chinese remainder theorem*, which will be very useful to us in Chapter 5. Let m_1 and m_2 be relatively prime and greater than 1. Given any two integers a and b, we show that we can always find a number x with $x \equiv a \pmod{m_1}$ and $x \equiv b \pmod{m_2}$. To satisfy the first requirement, write $x = qm_1 + a$ for some q to be determined. We want $qm_1 + a \equiv b \pmod{m_2}$. Since $\gcd(m_1, m_2) = 1$, by Lemma 3.9 there is some u with $m_1 u \equiv 1 \pmod{m_2}$. Multiplying by u gives the equivalent condition $q + au \equiv bu \pmod{m_2}$. Thus setting $q = (b-a)u + tm_2$ will give us a solution x for each integer t. More generally, we have the following.

Theorem 3.12 (Chinese remainder theorem)**.** *Let $m_1, m_2, \ldots, m_k$ be pairwise relatively prime numbers, each greater than 1. For any integers $b_1, b_2, \ldots, b_k$ there is an integer x satisfying $x \equiv b_i \pmod{m_i}$ for each $1 \leq i \leq k$.*

Proof. Let $N = m_1 m_2 \cdots m_k$, and put $n_i = N/m_i$. Then we have $\gcd(n_i, m_i) = 1$, and so by Theorem 3.2 there are integers x_i and y_i with $n_i x_i + m_i y_i = 1$. Set $e_i = n_i x_i$. Then $e_i \equiv 1 \pmod{m_i}$ and $e_i \equiv 0 \pmod{m_j}$ for $j \neq i$ since n_i is divisible by m_j for $j \neq i$.

Putting $x = b_1e_1 + b_2e_2 + \cdots + b_ke_k$ then satisfies the system of congruences. □

3.2. The Sum of Two Squares

We examine the equation

$$x^2 \equiv -1 \pmod{p}, \tag{3.1}$$

when p is of the form $4k + 3$. Clearly, 0 is not a solution to the congruence. If there is a nonzero solution, we can raise both sides to the power $(p-1)/2 = 2k+1$, which is an odd number. The left-hand side becomes x^{p-1}, which is congruent to 1 modulo p by Fermat's little theorem (Theorem 3.10). Since $1 \equiv -1 \pmod{p}$ only for $p = 2$, which is not of the form $4k + 3$, this is a contradiction. Thus the congruence has no solutions.

What about primes of the form $4k + 1$? Does (3.1) have a solution for these primes? We will answer this using Wilson's theorem (Theorem 3.11). We pair 1 with $p-1$, 2 with $p-2$, and so on, up to $(p-1)/2$ with $p-(p-1)/2 = (p+1)/2$. Since $p-k \equiv -k \pmod{p}$, we see that

$$(p-1)! \equiv (-1)^{(p-1)/2}[(p-1)/2]!^2 \pmod{p}.$$

Wilson's theorem yields $(p-1)! \equiv -1 \pmod{p}$. Since $(p-1)/2 = 2k$ is even, we have $[(p-1)/2]!^2 \equiv -1 \pmod{p}$. Thus $x = [(p-1)/2]!$ is a solution to (3.1).

We have proved the following theorem.

Theorem 3.13. *Let p be an odd prime. The congruence*

$$x^2 \equiv -1 \pmod{p}$$

has solutions if and only if $p \equiv 1 \pmod{4}$.

In the remainder of this section, we will determine which numbers can be written as the sum of two squares. To begin, we note the matrix product

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix}\begin{pmatrix} c & d \\ -d & c \end{pmatrix} = \begin{pmatrix} ac-bd & ad+bc \\ -(ad+bc) & ac-bd \end{pmatrix}.$$

Taking determinants yields the identity

$$(a^2+b^2)(c^2+d^2) = (ac-bd)^2 + (ad+bc)^2. \tag{3.2}$$

Thus if two numbers can both be written as a sum of two squares, then their product can also be written as a sum of two squares. To determine which numbers can be written as the sum of two squares, we must determine which prime numbers can be written as the sum of two squares. Identity (3.2) above is a special case of *Brahmagupta's identity*, which we will encounter in our study of the Brahmagupta–Pell equation in Section 3.4.

Let us observe the matrix identity used above is really an identity for the multiplication of two complex numbers. Indeed, if $i = \sqrt{-1}$ and we have $z = a + bi$, then $\overline{z} = a - bi$. We define $|z|^2 = z\overline{z}$. We see that $|z|^2 = z\overline{z} = a^2+b^2$. If $w = c+di$, then $zw = (ac-bd)+(ad+bc)i$ so that

$$|zw|^2 = zw\overline{zw} = |z|^2|w|^2,$$

and this is really the identity used above.

Now let us consider which primes can be written as a sum of two squares. Clearly, 2 can be written as such. If p is of the form $4k+3$ and

$$p = x^2 + y^2,$$

then reducing modulo 4 gives us a solution to $x^2 + y^2 \equiv 3 \pmod 4$, which is impossible since squares are congruent to either 0 or 1 modulo 4. Thus primes of this form cannot be written as a sum of two squares.

We will prove that every prime $p \equiv 1 \pmod 4$ can be written as a sum of two squares. Fix such a p, and consider the set

$$\{m : x^2 + y^2 = mp \text{ for some } x, y\}.$$

This set consists of those m for which mp can be written as a sum of two squares. By virtue of Theorem 3.13, $x^2 + 1 \equiv 0 \pmod p$ has a solution and so this set is not empty. Choose a minimal element m_0. We want to show that $m_0 = 1$. Suppose that $m_0 \geq 2$. Then we have

$$x^2 + y^2 = m_0 p,$$

for some x, y. Choose x_0 and y_0 satisfying $x \equiv x_0 \pmod{m_0}$ and $y \equiv y_0 \pmod{m_0}$, with $|x_0|, |y_0| \leq m_0/2$. Then

$$x_0^2 + y_0^2 \equiv x^2 + y^2 \equiv m_0 p \equiv 0 \pmod{m_0}.$$

Thus

$$x_0^2 + y_0^2 = m_0 m_1$$

for some m_1. Moreover, $|m_0 m_1| = |x_0^2 + y_0^2| \leq |x_0|^2 + |y_0|^2 \leq m_0^2/2$, so that $|m_1| \leq m_0/2$. Applying Brahmagupta's identity (3.2) with $(a, b) = (x, y)$ and $(c, d) = (x_0, -y_0)$, we obtain

$$(xx_0 + yy_0)^2 + (-xy_0 + yx_0)^2 = m_0^2 m_1 p.$$

We have $xx_0 + yy_0 \equiv x_0^2 + y_0^2 \equiv 0 \pmod{m_0}$ and consequently $xx_0 + yy_0$ is divisible by m_0. The second term $-xy_0 + yx_0$ satisfies $-xy_0 + yx_0 \equiv -x_0 y_0 + y_0 x_0 \equiv 0 \pmod{m_0}$, so it too is divisible by m_0. Thus, we have

$$\left(\frac{xx_0 + yy_0}{m_0}\right)^2 + \left(\frac{-xy_0 + yx_0}{m_0}\right)^2 = m_1 p.$$

That is, $m_1 p$ can also be written as a sum of two integral squares. This is a contradiction since $m_1 \leq m_0/2 < m_0$. Thus $m_0 = 1$ and p can be so written as a sum of two squares.

We can now determine which natural numbers can be written as the sum of two squares. The above discussion shows which primes can be so represented. If n can be represented as a sum of two squares and p is a prime dividing n with $p \equiv 3 \pmod{4}$, then we claim that it must appear to an even power. For suppose for such a p we have $p \mid n$. Reducing $x^2 + y^2 = n$ modulo p yields

$$x^2 \equiv -y^2 \pmod{p}.$$

Suppose p divides neither x nor y. Then the left-hand side is a square, but Theorem 3.13 implies the right-hand side cannot be. This is a contradiction. Thus one of x or y is divisible by p, and hence both must be. This yields $(x/p)^2 + (y/p)^2 = n/p^2$ for integers x/p, y/p, and n/p^2. By induction, we may deduce that the greatest power of p that divides n is even.

Conversely, suppose

$$n = 2^\alpha \prod_{p \equiv 1 \pmod{4}} p^{\beta_p} \prod_{q \equiv 3 \pmod{4}} q^{2\gamma_p},$$

where the products run over primes p and q. Since 2 and primes congruent to 1 modulo 4 can be written as a sum of two squares, so can their products. Observe that the product over primes congruent to 3 modulo 4 is a square. Since $m^2(x^2+y^2) = (mx)^2 + (my)^2$, it follows that n can be written as the sum of two squares.

To summarize, we have proved the following theorem.

Theorem 3.14. *A number may be written as the sum of two squares if and only if it has the form*

$$2^\alpha \prod_{p \equiv 1 \pmod 4} p^{\beta_p} \prod_{q \equiv 3 \pmod 4} q^{2\gamma_p},$$

where the products run over primes p and q.

3.3. The Sum of Four Squares

In the following theorem, we use a method similar to the previous section to determine which numbers may be written as the sum of four squares.

Theorem 3.15. *Every natural number can be written as a sum of four squares.*

This is a theorem of Joseph-Louis Lagrange (1736–1813), who proved it in 1770. We prove it in four steps.

Consider the following matrix identity for complex numbers:

$$\begin{pmatrix} z & w \\ -\overline{w} & \overline{z} \end{pmatrix} \begin{pmatrix} u & v \\ -\overline{v} & \overline{u} \end{pmatrix} = \begin{pmatrix} uz - w\overline{v} & zv + w\overline{u} \\ -\overline{w}u - \overline{v}\overline{z} & \overline{u}\overline{z} - \overline{w}v \end{pmatrix}.$$

Taking determinants, we obtain the identity

$$(|z|^2 + |w|^2)(|u|^2 + |v|^2) = |uz - \overline{v}w|^2 + |w\overline{u} + zv|^2.$$

This is also easy to verify directly for all complex numbers u, v, w, z. We deduce from it, by putting $z = x_1 + ix_2$, $w = x_3 + ix_4$, $u = y_1 + iy_2$, and $v = y_3 + iy_4$, that

$$(x_1^2 + x_2^2 + x_3^2 + x_4^2)(y_1^2 + y_2^2 + y_3^2 + y_4^2) = z_1^2 + z_2^2 + z_3^2 + z_4^2,$$

where

$$z_1 = x_1y_1 - x_2y_2 - x_3y_3 - x_4y_4,$$
$$z_2 = x_1y_2 + x_2y_1 + x_3y_4 - x_4y_3,$$

$$z_3 = x_1y_3 - x_2y_4 + x_3y_1 + x_4y_2,$$
$$z_4 = x_1y_4 + x_2y_3 - x_3y_2 + x_4y_1.$$

Thus, if two numbers can both be written as a sum of four squares, then their product also can be written in this way. Furthermore, we have an explicit recipe for determining the squares to sum for the product, given the squares summing to each factor. As every number is a product of prime numbers, it therefore suffices to prove Lagrange's theorem for prime numbers. Since $2 = 1^2+1^2+0^2+0^2$, we need only focus on odd primes.

The next step is to see that for any odd prime p, we can solve the congruence

$$x^2 + y^2 + 1 \equiv 0 \pmod{p}.$$

To see this, we consider the set of squares modulo p, which has size[7] $(p+1)/2$. The same is true of the set of elements of the form $-1-y^2$. If these sets were disjoint, we would have at least $p+1$ residue classes modulo p, a contradiction. Hence there is a common element, which gives a solution to the congruence. Since the integers $-(p-1)/2,\ldots,(p-1)/2$ form a complete set of residue classes modulo p, we may take $|x| < p/2$ and $|y| < p/2$. We deduce that there are integers x and y so that

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

with $m < p$.

The third step is to consider the smallest positive integer m such that mp can be written as a sum of four squares. By the previous paragraph, the set of all such m is nonempty. Let m_0 be the smallest such m. Then $m_0 < p$, again by the previous paragraph. If $m_0 = 1$, we are done, so let us suppose that $1 < m_0 < p$. Hence we can write

$$m_0p = x_1^2 + x_2^2 + x_3^2 + x_4^2. \tag{3.3}$$

Suppose m_0 is even. Considering (3.3) modulo 2, we see that there are three possibilities: that each x_i is even; that each is odd; or that

[7] Since $(p-k)^2 \equiv k^2 \pmod{p}$, all squares can be found in $S = \{0^2,\ldots,(\frac{p-1}{2})^2\}$. Suppose $i^2 \equiv j^2 \pmod{p}$ for $0 \le i < j \le \frac{p-1}{2}$. Then $p \mid (j-i)(j+i)$, and so $p \mid (j-i)$ or $p \mid (j+i)$. Our bounds on i and j yield $0 < j-i \le \frac{p-1}{2}$ and $0 < j+i < p-1$, and so neither $i-j$ nor $i+j$ can be divisible by p. Thus S contains all squares modulo p exactly once.

precisely two of them, without loss of generality, say x_1 and x_2, are even. In any of these cases, x_1+x_2, x_1-x_2, x_3+x_4, and x_3-x_4 are even. The identity

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

then yields

$$\frac{m_0}{2}p = \left(\frac{x_1+x_2}{2}\right)^2 + \left(\frac{x_1-x_2}{2}\right)^2 + \left(\frac{x_3+x_4}{2}\right)^2 + \left(\frac{x_3-x_4}{2}\right)^2.$$

Thus $(m_0/2)p$ can be written as a sum of four squares, which contradicts the minimality of m_0. Hence we may suppose m_0 is odd.

The final step involves choosing y_1, y_2, y_3, and y_4 so that $y_1 \equiv x_1 \pmod{m_0}$ and $y_i \equiv -x_i \pmod{m_0}$ for $2 \le i \le 4$, with each $|y_i| \le (m_0-1)/2$. Then (3.3) yields

$$m_0 m_1 = y_1^2 + y_2^2 + y_3^2 + y_4^2$$

with $0 < m_1 < m_0$. Since both $m_0 p$ and $m_0 m_1$ can be written as a sum of four squares, so can their product:

$$(m_0 p)(m_0 m_1) = z_1^2 + z_2^2 + z_3^2 + z_4^2,$$

with each z_i given explicitly in terms of the x_i and y_i. From this explicit description, we see directly that each $z_i \equiv 0 \pmod{m_0}$. For example, we have

$$\begin{aligned} z_1 &= x_1y_1 - x_2y_2 - x_3y_3 - x_4y_4 \\ &\equiv x_1^2 + x_2^2 + x_3^2 + x_4^2 \pmod{m_0} \\ &\equiv 0 \pmod{m_0}. \end{aligned}$$

Thus we may divide out by m_0^2 and deduce that $m_1 p$ can be written as a sum of four squares. But this contradicts the minimality of m_0 as $m_1 < m_0$. Hence, we must have $m_0 = 1$ in (3.3), completing the proof of Lagrange's theorem.

3.4. The Brahmagupta–Pell Equation

We will discuss the equation

$$x^2 - dy^2 = 1$$

for $d \in \mathbb{Z}$. This equation was first studied in 6th century CE by the Indian mathematician Brahmagupta. He discovered the *chakravala* method of generating many solutions from one solution. The word *chakra* in Sanskrit means "wheel" indicating that Brahmagupta was aware of the cyclic nature of the set of all solutions. Indeed, in modern parlance, the set of solutions can be given the structure of an infinite cyclic group. Over the centuries, several Indian mathematicians refined the techniques of Brahmagupta, notably Jayadeva and Bhaskaracharya in the 12th century. They essentially discovered the continued fraction algorithm for finding the *minimal solution* of the equation from which all the solutions can be generated by the *chakravala* method. The equation was rediscovered by Pierre de Fermat (1607–1665), but Leonhard Euler (1707–1783) incorrectly attributed it to John Pell (1611–1685), and since then the name has stuck. Even though the Indian mathematicians of antiquity had the algorithm for finding all the solutions, they had not written down the proof that their method gives all the solutions. Perhaps such a refined notion of proof was not extant at that time. Nevertheless, the completion of this finer detail was given by Lagrange. We refer the read to Weil's excellent historical account [**We**].

When discussing solutions to the equation $x^2 - dy^2 = 1$, we can confine ourselves to natural number solutions, since (x, y) is a solution implies that $(\pm x, \pm y)$ is also a solution. We note the trivial solution $(1, 0)$. For $d \leq -2$, both x^2 and $-dy^2$ are positive. We must take $y = 0$ if they are to sum to 1, leaving us with only the trivial solution. For $d = -1$, we have $(0, 1)$ in addition to the trivial solution. For $d = 0$, $(\pm 1, y)$ is a solution for any y. If $d = n^2$ is a perfect square, then $1 = x^2 - dy^2 = (x - ny)(x + ny)$. This implies that $x - ny$ and $x + ny$ are either both 1 or both -1, from which it is easily deduced that the only solution is the trivial solution. Thus in what follows, we will be taking $d \geq 2$ not a square.

Following our earlier approach using matrices, one can write

$$\begin{pmatrix} x_0 & y_0 \\ dy_0 & x_0 \end{pmatrix} \begin{pmatrix} x_1 & y_1 \\ dy_1 & x_1 \end{pmatrix} = \begin{pmatrix} x_0x_1 + dy_0y_1 & x_0y_1 + y_0x_1 \\ d(x_0y_1 + y_0x_1) & x_0x_1 + dy_0y_1 \end{pmatrix}.$$

Taking determinants yields Brahmagupta's identity

$$(x_0^2 - dy_0^2)(x_1^2 - dy_1^2) = (x_0x_1 + dy_0y_1)^2 - d(x_0y_1 + y_0x_1)^2.$$

Thus if (x_0, y_0) and (x_1, y_1) satisfy $x^2 - dy^2 = 1$, then so does

$$(x_0x_1 + dy_0y_1, x_0y_1 + y_0x_1).$$

In other words, we can make new solutions from old solutions.

As an example, consider the equation $x^2 - 2y^2 = 1$. We can start looking for a solution by running through values of y and testing if $2y^2 + 1$ is a square. Doing this, we see that $(3, 2)$ is a nontrivial solution. Using the above identity on $(3, 2)$ with itself yields $(17, 12)$ as another solution. In fact, it turns out there are no solutions with a y-value between 2 and 12. Using the identity again on solutions $(3, 2)$ and $(17, 12)$ yields another solution $(99, 70)$. We can repeat this process as much as we desire.

Thus we see that *if* we can find a nontrivial solution to the Brahmagupta–Pell equation, we can then use the above identity to generate infinitely many solutions. In this section we will show that a nontrivial solution always exists, and that *all* solutions to the equation are generated using this method.

To each integer solution (x, y) of $x^2 - dy^2 = 1$, we attach the real number $r = x + y\sqrt{d}$ and say r *represents* the solution (x, y). In our above example, $3 + 2\sqrt{2}$ represents the solution $(3, 2)$. For an integer solution (x, y) we have

$$\frac{1}{x + y\sqrt{d}} = \frac{x - y\sqrt{d}}{(x + y\sqrt{d})(x - y\sqrt{d})} = \frac{x - y\sqrt{d}}{x^2 - dy^2} = x - y\sqrt{d}.$$

Thus if r represents an integer solution, then so does $1/r$. For integer solutions (x_0, y_0) and (x_1, y_1) we have

$$(x_0 + y_0\sqrt{d})(x_1 + y_1\sqrt{d}) = (x_0x_1 + dy_0y_1) + (x_0y_1 + y_0x_1)\sqrt{d}.$$

Since Brahmagupta's identity above implies $(x_0x_1 + dy_0y_1, x_0y_1 + y_0x_1)$ is a solution, it follows that if r and s represent solutions, then so does rs. We are now ready to prove the following lemma.

Lemma 3.16. *If r represents an integer solution of the Brahmagupta–Pell equation, then r^k will also represent an integer solution for any integer k.*

Proof. Setting $k = 0$ yields the trivial solution. We may use induction to show that the result holds for positive k, since if r and r^k represent solutions, then so does the product $rr^k = r^{k+1}$. If k is negative, then $-k$ is positive, and hence r^{-k} represents a solution. It follows that the reciprocal $1/r^{-k} = r^k$ also represents a solution. □

The solutions found in our above example with $d = 2$ can now be more compactly described by the equalities $(3 + 2\sqrt{2})^2 = 17 + 12\sqrt{2}$ and $(3 + 2\sqrt{2})^3 = 99 + 70\sqrt{2}$.

We may order the natural number solutions of the Brahmagupta–Pell equation according to their representatives.

Lemma 3.17. *Let (x, y) and (w, z) be natural number solutions to the Brahmagupta–Pell equation with representatives r and s, respectively. Then $r < s$ if and only if $y < z$.*

Proof. To prove the forward direction, suppose $y \geq z$. Then $x^2 = 1 + dy^2 \geq 1 + dz^2 = w^2$. Since x and w are nonnegative, this implies $x \geq w$. Since $y \geq z$ and $x \geq w$, we have $r = x + y\sqrt{d} \geq w + z\sqrt{d} = s$. The proof of the reverse direction is almost identical, only with $<$ replacing $\geq$. □

At the end of this section, we will show that the Brahmagupta–Pell equation (with $d \geq 2$ not a square) always has a nontrivial solution. For now, we will assume this result.

Theorem 3.18. *Let α be the least representative of a nontrivial natural number solution (x, y) to the Brahmagupta–Pell equation. Then all nontrivial natural number solutions of $x^2 - dy^2 = 1$ are given by α^k for some $k \geq 1$. We call α the generator for $x^2 - dy^2 = 1$.*

Proof. The representatives for natural number solutions are not integers, but by Lemma 3.17, they are ordered by the second coordinates of their corresponding solutions, which are nonnegative integers. Hence this minimum α will always exist. Since the representative of a nontrivial natural number solution must be greater than 1, we have $\alpha > 1$. By Lemma 3.16, α^k represents an integer solution, but since $\alpha > 1$, it will in fact be a natural number solution. Now let (x, y) be a nontrivial natural number solution with representative r, so that

$r \geq \alpha > 1$. This implies the existence of $k \in \mathbb{N}$ with $\alpha^k \leq r < \alpha^{k+1}$, and so $1 \leq r\alpha^{-k} < \alpha$. By Lemma 3.16, α^{-k} represents a solution. Since α is the smallest representative greater than 1 that yields a nontrivial solution, we must have $r\alpha^{-k} = 1$. Hence $r = \alpha^k$, as required. □

We can use the solutions of the Brahmagupta–Pell equation to give quickly converging rational approximations to $\sqrt{d}$. Rearranging $x^2 - dy^2 = 1$, we have $\sqrt{d} = \frac{\sqrt{x^2-1}}{y}$, which is very close to $\frac{x}{y}$. For example, when $d = 2$, we have generator $3 + 2\sqrt{2}$. By successively squaring, we can easily compute (by hand, even)

$$\begin{aligned}
&(3 + 2\sqrt{2})^2 = 17 + 12\sqrt{2},\\
&(3 + 2\sqrt{2})^4 = (17 + 12\sqrt{2})^2 = 577 + 408\sqrt{2}, \text{ and}\\
&(3 + 2\sqrt{2})^8 = (577 + 408\sqrt{2})^2 = 665857 + 470832\sqrt{2}.
\end{aligned}$$

Thus 665857/470832 should be close to $\sqrt{2}$, and indeed it is accurate to 11 decimal places. One further squaring yields

$$886731088897/627013566048,$$

accurate to 23 decimal places!

Letting (x_1, y_1) be the solution represented by generator α, we define natural numbers x_k and y_k by

$$x_k + y_k\sqrt{d} = \left(x_1 + y_1\sqrt{d}\right)^k. \tag{3.4}$$

By our work above, $\{(x_k, y_k) : k \in \mathbb{N}\}$ contains all the natural number solutions of $x^2 - dy^2 = 1$. Note that

$$\begin{aligned}
x_k - y_k\sqrt{d} &= \frac{(x_k - y_k\sqrt{d})(x_k + y_k\sqrt{d})}{x_k + y_k\sqrt{d}}\\
&= \frac{1}{(x_1 + y_1\sqrt{d})^k}\\
&= \left(\frac{x_1 - y_1\sqrt{d}}{(x_1 + y_1\sqrt{d})(x_1 - y_1\sqrt{d})}\right)^k\\
&= (x_1 - y_1\sqrt{d})^k.
\end{aligned} \tag{3.5}$$

Adding this to (3.4) and dividing by 2 yields

$$x_k = \frac{1}{2}\left((x_1 + y_1\sqrt{d})^k + (x_1 - y_1\sqrt{d})^k\right).$$

Subtracting (3.5) from (3.4) and dividing by $2\sqrt{d}$ yields

$$y_k = \frac{1}{2\sqrt{d}}\left((x_1 + y_1\sqrt{d})^k - (x_1 - y_1\sqrt{d})^k\right).$$

We note that $x_1 - y_1\sqrt{d} = \frac{1}{x_1+y_1\sqrt{d}}$, and hence $0 < x_1 - y_1\sqrt{d} < 1$. Thus $0 < \frac{1}{2}(x_1 - y_1\sqrt{d})^k < \frac{1}{2}$, and so x_k is the integer closest to $\frac{1}{2}(x_1+y_1\sqrt{d})^k$. Similarly, y_k is the closest integer to $\frac{1}{2\sqrt{d}}(x_1+y_1\sqrt{d})^k$. Thus the solutions to the Brahmagupta–Pell equation are growing exponentially.

We now give some results on the x_k and y_k.

Theorem 3.19. *The following addition and subtraction formulas hold:*

$$\begin{aligned} x_{k+\ell} &= x_k x_\ell + d y_k y_\ell, \\ y_{k+\ell} &= x_k y_\ell + x_\ell y_k, \\ x_{k-\ell} &= x_k x_\ell - d y_k y_\ell, \\ y_{k-\ell} &= x_\ell y_k - x_k y_\ell. \end{aligned}$$

Proof. Using defining relation $x_k + y_k\sqrt{d} = \left(x_1 + y_1\sqrt{d}\right)^k$, we have

$$\begin{aligned} x_{k+\ell} + y_{k+\ell}\sqrt{d} &= \left(x_1 + y_1\sqrt{d}\right)^{k+\ell} \\ &= \left(x_1 + y_1\sqrt{d}\right)^k \left(x_1 + y_1\sqrt{d}\right)^\ell \\ &= \left(x_k + y_k\sqrt{d}\right)\left(x_\ell + y_\ell\sqrt{d}\right) \\ &= (x_k x_\ell + d y_k y_\ell) + (x_k y_\ell + x_\ell y_k)\sqrt{d}. \end{aligned}$$

This yields the addition formulas. Similarly, we have

$$\left(x_{k-\ell} + y_{k-\ell}\sqrt{d}\right)\left(x_\ell + y_\ell\sqrt{d}\right) = x_k + y_k\sqrt{d}.$$

Multiplying both sides by $x_\ell - y_\ell\sqrt{d}$ yields

$$\begin{aligned} x_{k-\ell} + y_{k-\ell}\sqrt{d} &= \left(x_k + y_k\sqrt{d}\right)\left(x_\ell - y_\ell\sqrt{d}\right) \\ &= (x_k x_\ell - d y_k y_\ell) + (x_\ell y_k - x_k y_\ell)\sqrt{d}, \end{aligned}$$

which yields the subtraction formulas. □

Theorem 3.20. *Let $(x_0, y_0) = (1, 0)$ be the trivial solution, and let (x_1, y_1) be the generator solution. Then the natural number solutions x_k and y_k are given by the recursive equations*

$$\begin{aligned} x_{k+1} &= 2x_1 x_k - x_{k-1}, \\ y_{k+1} &= 2x_1 y_k - y_{k-1}. \end{aligned}$$

Proof. To get the first equation, let $\ell = 1$ and add the formulas for x_{k+1} and x_{k-1} of Theorem 3.19. Adding the formulas for y_{k+1} and y_{k-1} from the same theorem yields the second recursive equation. □

The recursive equations of Theorem 3.20 allow us to use induction to prove results on the sequences x_k and y_k.

Lemma 3.21. *The sequences x_k and y_k are increasing.*

Proof. First off, we note that $x_1 \geq 2$, since $x_1 = 0$ is impossible and $x_1 = 1$ yields the trivial solution. Thus $x_1 > 1 = x_0$. Suppose $x_k - x_{k-1} > 0$. Then

$$\begin{aligned} x_{k+1} - x_k &= 2x_1 x_k - x_{k-1} - x_k \\ &\geq 3x_k - x_{k-1} \\ &> x_k - x_{k-1} \\ &> 0. \end{aligned}$$

Thus by induction, the x_k are increasing. The proof that the y_k are increasing is similar. □

We now give some bounds on x_k and y_k.

Lemma 3.22. *We have*

$$x_1^k \leq x_k \leq (2x_1)^k$$

and

$$k \leq y_k \leq (2x_1)^k.$$

Proof. We use induction to show the first bound. Since $x_0 = 1$, the case $k = 0$ is clear. Suppose the result holds for k. The addition formula of Theorem 3.19 with $\ell = 1$ yields

$$\begin{aligned} x_{k+1} &= x_k x_1 + d y_k y_1 \\ &\geq x_k x_1 \\ &\geq x_1^{k+1}. \end{aligned}$$

The recursive formula of Theorem 3.20 yields

$$\begin{aligned} x_{k+1} &= 2x_1 x_k - x_{k-1} \\ &\leq 2x_1 x_k \\ &\leq (2x_1)^{k+1}. \end{aligned}$$

We now show the second bound, again with induction. Since $y_0 = 0$, the case $k = 0$ is clear. Suppose the result holds for k. Since by Theorem 3.21 the y_k are increasing, it follows that $y_{k+1} > y_k \geq k$. Thus $y_{k+1} \geq k + 1$, as required. The recursive formula of Theorem 3.20 with $\ell = 1$ yields

$$\begin{aligned} y_{k+1} &= 2x_1 y_k - y_{k-1} \\ &\leq 2x_1 y_k \\ &< (2x_1)^{k+1}. \end{aligned}$$

This completes the proof. □

We prove a divisibility result for the y_k.

Theorem 3.23. *$y_k \mid y_\ell$ if and only if $k \mid \ell$.*

Proof. First, we note that $\gcd(x_k, y_k) = 1$. To see this, suppose $p \mid x_k$ and $p \mid y_k$. This implies $p \mid (x_k^2 - d y_k^2)$, and hence $p \mid 1$. We now show that

$$y_k \mid y_\ell \text{ if and only if } k \mid \ell.$$

We begin with the reverse direction. Suppose $k \mid \ell$, so that $\ell = nk$ for $n \in \mathbb{N}$. We show that $y_k \mid y_{nk}$ with induction on n. The result is clear when $n = 1$. By the addition formula of Theorem 3.19, we have

$$y_{(n+1)k} = x_{nk} y_k + x_k y_{nk}.$$

Thus if $y_k \mid y_{nk}$, then $y_k \mid y_{(n+1)k}$. We now show the forward direction. Suppose $y_k \mid y_\ell$, and write $\ell = qk + r$ for $q, r \in \mathbb{N}$, $0 \leq r < k$. By the addition formula of Theorem 3.19, we have

$$y_\ell = x_{qk} y_r + x_r y_{qk}.$$

By our work in the reverse direction above, $y_k \mid y_{qk}$ and hence $y_k \mid x_r y_{qk}$. Since we also have $y_k \mid y_\ell$, it follows that $y_k \mid x_{qk} y_r$. We claim that $\gcd(y_k, x_{qk}) = 1$, for suppose $p \mid y_k$ and $p \mid x_{qk}$. Again by our work in the reverse direction, it follows that $y_k \mid y_{qk}$. Since $p \mid y_k$, we have $p \mid y_{qk}$. Since $\gcd(x_{qk}, y_{qk}) = 1$, it follows that $p = 1$. Hence we must have $y_k \mid y_r$ with $0 \leq r < k$. Since the y_k are an increasing sequence of natural numbers, it must be that $y_r = 0$, and hence $r = 0$. Thus $\ell = qk$, and so $k \mid \ell$. □

In the remainder of this section, we prove the existence of a nontrivial solution to the Brahmagupta–Pell equation for $d \geq 2$ not a square. To begin with, we prove the *Dirichlet approximation theorem*. This theorem was proved by Peter Gustav Lejeune Dirichlet (1805–1859).

Theorem 3.24 (Dirichlet approximation theorem)**.** *Given a real number α and any positive integer N, we can find integers p and q with $1 \leq q \leq N$ such that $|p - q\alpha| < \frac{1}{N}$.*

Proof. First, we note that we need only prove the theorem for positive α, since $|p - q(-\alpha)| = |-p - q\alpha|$. Let $\{x\}$ denote the *fractional part* of x, so that $\{x\} = x - \lfloor x \rfloor$, where $\lfloor x \rfloor$ is the greatest integer less than or equal to x. Thus, for example, $\{2.997\} = 0.997$ and $\{5\} = 0$. Then

$$0, \{\alpha\}, \{2\alpha\}, \ldots, \{N\alpha\}$$

are $N + 1$ nonnegative numbers each less than 1. The N sets

$$\left\{ x : \frac{m}{N} \leq x < \frac{m+1}{N} \right\}$$

for $m = 0, \ldots, N - 1$ form a partition of the interval $\{x : 0 \leq x < 1\}$. Since we have $N + 1$ numbers falling in N subintervals, it must be the case that there is some subinterval containing at least two of

the numbers.[8] Since the distance between any two members of a subinterval is less than $\frac{1}{N}$, there exist $0 \le i < j \le N$ with

$$0 \le |\{i\alpha\} - \{j\alpha\}| < \frac{1}{N}. \tag{3.6}$$

However, we have

$$\begin{aligned}\{i\alpha\} - \{j\alpha\} &= i\alpha - \lfloor i\alpha \rfloor - (j\alpha - \lfloor j\alpha \rfloor) \\ &= \lfloor j\alpha \rfloor - \lfloor i\alpha \rfloor - (j-i)\alpha.\end{aligned}$$

We take $q = j - i$ and $p = \lfloor j\alpha \rfloor - \lfloor i\alpha \rfloor$, both integers. Then (3.6) becomes

$$0 \le |p - q\alpha| < \frac{1}{N}.$$

Since $0 \le i < j \le N$, we have $0 < j - i \le N$, and so $1 \le q \le N$. □

Lemma 3.25. *For $d \ge 2$ not a square, the inequality*

$$\left| x - y\sqrt{d} \right| < \frac{1}{y}$$

has infinitely many positive solutions.

Proof. Setting $x = \lfloor \sqrt{d} \rfloor$, $y = 1$ yields one solution of the inequality

$$\left| x - y\sqrt{d} \right| < \frac{1}{y}. \tag{3.7}$$

Given a solution (x_1, y_1) of (3.7), we may take N large enough so that

$$\frac{1}{N} < \left| x_1 - y_1\sqrt{d} \right|,$$

as the right-hand side cannot be equal to zero since $\sqrt{d}$ is irrational. The Dirichlet approximation theorem then yields (x_2, y_2) with

$$\left| x_2 - y_2\sqrt{d} \right| < \frac{1}{N} < \left| x_1 - y_1\sqrt{d} \right|$$

and $1 \le y_2 \le N$. Thus $\frac{1}{N} \le \frac{1}{y_2}$, and hence (x_2, y_2) is also a solution of (3.7). We note that (3.7) implies $-\frac{1}{y} < x - y\sqrt{d}$, and hence

$$x_2 > y_2\sqrt{d} - \frac{1}{y_2} \ge \sqrt{d} - 1 > 0.$$

[8] We have made use of the *pigeonhole principle* here. The pigeonhole principle states that if $n + 1$ or more objects are placed in n boxes, then at least one box will contain two or more objects. Although this principle sounds fairly obvious at first encounter, it has many surprisingly deep consequences!

Thus x_2 and y_2 are positive solutions to (3.7). In this way, we may create an infinite sequence of positive solutions (x_n, y_n) to (3.7). Since $|x_n - y_n\sqrt{d}|$ is decreasing, the solutions are distinct. □

We are now ready to prove the existence of a nontrivial solution to the Brahmagupta–Pell equation.

Theorem 3.26. *When $d \geq 2$ is not a square, the Brahmagupta–Pell equation*

$$x^2 - dy^2 = 1$$

has a nontrivial solution.

Proof. Consider a solution to

$$\left|x - y\sqrt{d}\right| < \frac{1}{y},$$

the inequality of Lemma 3.25. For such a solution, we have

$$\begin{aligned}\left|x + y\sqrt{d}\right| &\leq \left|x - y\sqrt{d}\right| + 2y\sqrt{d} \\ &< \frac{1}{y} + 2y\sqrt{d} \\ &\leq y + 2y\sqrt{d} \\ &= \left(1 + 2\sqrt{d}\right) y.\end{aligned}$$

Thus

$$\begin{aligned}\left|x^2 - dy^2\right| &= \left|x + y\sqrt{d}\right| \left|x - y\sqrt{d}\right| \\ &< \left(1 + 2\sqrt{d}\right) y\frac{1}{y} \\ &= 1 + 2\sqrt{d}.\end{aligned}$$

Note that since $\sqrt{d}$ is irrational, $|x^2 - dy^2| > 0$. Since the inequality of Lemma 3.25 has infinitely many solutions and there are only finitely many natural numbers between 0 and $1 + 2\sqrt{d}$, there is some integer k with $0 < |k| < 1 + 2\sqrt{d}$ for which $x^2 - dy^2 = k$ has infinitely many positive solutions.[9]

[9] We have used an infinite version of the pigeonhole principle here: if infinitely many objects are placed in a finite number of boxes, then at least one box will contain infinitely many objects.

Now let k be such that $x^2 - dy^2 = k$ has infinitely many positive solutions. For a solution (x, y), x is congruent to one of $0, 1, \ldots, |k|-1$ modulo $|k|$, as is y. There are k^2 possible pairs of residue classes for x and y. Since there are infinitely many solutions, at least one pair of residue classes contains infinitely many and, hence, two solutions. Let us name these positive solutions (x_1, y_1) and (x_2, y_2). Then $(x_1, y_1) \neq (x_2, y_2)$ with $x_1 \equiv x_2 \pmod{|k|}$ and $y_1 \equiv y_2 \pmod{|k|}$. We set

$$x = \frac{x_1x_2 - dy_1y_2}{k} \quad \text{and} \quad y = \frac{x_1y_2 - x_2y_1}{k}.$$

Since

$$x_1x_2 - dy_1y_2 \equiv x_1^2 - dy_1^2 \equiv k \equiv 0 \pmod{|k|}$$

and

$$x_1y_2 - x_2y_1 \equiv x_1y_1 - x_1y_1 \equiv 0 \pmod{|k|},$$

x and y are integers. We have

$$\begin{aligned} x^2 - dy^2 &= \left(\frac{x_1x_2 - dy_1y_2}{k}\right)^2 - d\left(\frac{x_1y_2 - x_2y_1}{k}\right)^2 \\ &= \frac{1}{k^2}\left(x_1^2x_2^2 + d^2y_1^2y_2^2 - dx_1^2y_2^2 - dx_2^2y_1^2\right) \\ &= \frac{1}{k^2}(x_1^2 - dy_1^2)(x_2^2 - dy_2^2) \\ &= 1, \end{aligned}$$

and so (x, y) is a solution to the Brahmagupta–Pell equation. It remains to show that this is not the trivial solution. Suppose $y = 0$ and $x = \pm 1$. Then $x_1y_2 = x_2y_1$ and $x_1x_2 - dy_1y_2 = \pm k$. Multiplying the second equation by y_2 and then using the first equation yields

$$\pm ky_2 = y_1\left(x_2^2 - dy_2^2\right) = ky_1.$$

Thus $y_1 = \pm y_2$. Since y_1 and y_2 are positive, we have $y_1 = y_2$. The first equation then implies $x_1 = x_2$, which contradicts the fact that $(x_1, y_1) \neq (x_2, y_2)$. Thus $(|x|, |y|)$ is a nontrivial natural number solution to the Brahmagupta–Pell equation. □

We note that although the argument above shows that a nontrivial solution of the Brahmagupta–Pell equation must exist, it does not give us a simple method to find the generator. The generator can get quite large. For example, the smallest nontrivial solution to

$x^2 - 61y^2 = 1$ is $x = 1766319049$, $y = 226153980$. The continued fraction of $\sqrt{d}$ can be used to find the generator. The reader can find a readable exposition in [**ME**, section 8.2].

Further Reading

There are many books that introduce the reader to elementary number theory. Alan Baker's aptly named book *A Concise Introduction to the Theory of Numbers* [**Ba1**] manages to cover a lot of ground in just over 100 pages. The book has been redeveloped into the larger *A Comprehensive Course in Number Theory* [**Ba2**]. Hardy and Wright's *An Introduction to the Theory of Numbers* [**HW**] is a classic introduction to the subject and has been updated to contain more recent material.

Exercises

3.1. Prove the following basic properties of division.
   (a) If $c \mid a$ and $c \mid b$, then $c \mid (a + b)$.
   (b) If $b \mid a$ and k is an integer, then $b \mid ka$.
   (c) If $b \mid a$ and $a \mid b$, then $a = \pm b$.

3.2. Use the Euclidean algorithm to find the greatest common divisor of 42823 and 6409.

3.3. Find all integers x and y such that $42823x + 6409y = 17$.

3.4. (a) Use the Euclidean algorithm to find $\gcd(78787, 11111)$.
   (b) Find integers s and t such that

$$11111s + 78787t = \gcd(78787, 11111).$$

   (c) Find all x such that

$$11111x \equiv 10 \pmod{78787}.$$

3.5. Let F_n be the nth Fibonacci number given by the initial values $F_1 = F_2 = 1$ and the recursion

$$F_{n+1} = F_n + F_{n-1}$$

for $n \geq 2$.

(a) Show that the greatest common divisor of F_n and F_{n+1} is 1.
(b) Show that the greatest common divisor of F_n and F_{n+2} is 1.
(c) Prove that $F_1 + F_2 + \cdots + F_{n-1} = F_{n+1} - 1$.
(d) Prove that $\gcd(F_n, F_m) = F_{\gcd(n,m)}$.

3.6. Let

$$S_n = \{m : 1 \leq m < n \text{ and } \gcd(m, n) = 1\},$$

and define $\phi(n) = |S_n|$. This is *Euler's totient function.* Let $a \in \mathbb{Z}$ with $\gcd(a, n) = 1$. Let T consist of natural numbers m with $1 \leq m < n$ and $m \equiv am_1 \pmod{n}$ for $m_1 \in S_n$. Show $T = S_n$. Since this implies that the product of all the elements of each set are congruent modulo n, deduce *Euler's theorem*:

$$a^{\phi(n)} \equiv 1 \pmod{n}.$$

3.7. Show that if n is a natural number dividing $2^n - 1$, then $n = 1$.

3.8. Let $d = \gcd(m, n)$. Show that for any natural number $a > 1$, we have

$$\gcd(a^m - 1, a^n - 1) = a^d - 1.$$

3.9. Find all ways in which 1000 be expressed as the sum of two positive integers, one of which is divisible by 11 and the other by 17.

3.10. Show that n is prime if and only if $(n-2)! \equiv 1 \pmod{n}$.

3.11. The earliest known use of the Chinese remainder theorem is from a problem in the *Sunzi Suanjing*, a Chinese work dating from the 3rd to 5th centuries CE. In it, the following problem is proposed and solved: "There are an unknown number of things. When counting by threes, two are left behind. When counting by fives, three are left. When counting by sevens, two are left. How many things are there?" Solve this problem.

3.12. Show that squares are congruent to 0, 1, or 4 modulo 8. Use this to deduce that if $n \equiv 7 \pmod{8}$, then n is not the sum of three squares.

3.13. Find the generator $x_1 + y_1\sqrt{5}$ to the Brahmagupta–Pell equation $x^2 - 5y^2 = 1$. Find $(x_1 + y_1\sqrt{5})^4$ by squaring twice, and use it

to give a rational approximation to $\sqrt{5}$ that is accurate to nine decimal digits. Squaring one more time will yield a rational approximation accurate to 19 decimal digits.

3.14. Complete the proof of Lemma 3.21 by using the recursive formula of Theorem 3.20 to show that y_k is an increasing sequence.

3.15. Suppose n is an integer. A set of natural numbers $\{a_1, \ldots, a_m\}$ is called a *Diophantine m-tuple*[10] *with property $D(n)$* if $a_i a_j + n$ is a perfect square for $1 \leq i < j \leq m$. If $n \equiv 2 \pmod 4$, show that $m \leq 3$. (It is conjectured that m is bounded for any n by an absolute constant, but this conjecture is still open.)

[10] Our use of the term "Diophantine m-tuple" here is different from the use of the term elsewhere in the text, particularly in Chapter 5.

Chapter 4

Computability and Provability

4.1. Turing Machines

We have now seen several examples of *algorithms*, which are, informally, effective computable procedures for solving a specific problem. They consist of an unambiguous set of instructions that can be carried out by a person with no innovation or originality required on their part. We will assume that this person has unlimited time and writing space for carrying out this procedure (which, in practice, is not often the case). We consider functions mapping from $\mathbb{N}$ to $\mathbb{N}$ or from n-tuples of $\mathbb{N}$ to $\mathbb{N}$, and call such a function *effectively computable* when we have an algorithm to calculate the value of the function when given any element of its domain. This is an informal definition but is typically sufficient when showing a function is computable; given a clear description of an algorithm, we can agree that it is indeed computing what is being claimed.

Any set of instructions consists of a finite string of letters and punctuation. In the exercises to Chapter 1, we saw that the number of finite strings of finitely many symbols is countably infinite, and hence there must be countably many effectively computable functions. We also saw that there are uncountably many functions from

$\mathbb{N}$ to $\mathbb{N}$, and so this implies that there are functions that are not effectively computable. In fact, in this sense, most functions are not effectively computable. However, many functions that are of mathematical interest are effectively computable.

A *Diophantine equation* is a polynomial equation with integral coefficients and any number of variables for which we are interested in finding integer solutions. The Brahmagupta–Pell equation of Section 3.4 is an example of a Diophantine equation.

Here is a question. Does the following system of Diophantine equations have any integer solutions?

$$\begin{aligned} 6w + 2x^2 - y^3 &= 0, \\ 5xy - z^2 + 6 &= 0, \\ w^2 - w + 2x - y + z - 4 &= 0. \end{aligned}$$

One can verify that $w = 1$, $x = 1$, $y = 2$, $z = 4$ is a solution. But now consider

$$\begin{aligned} 6w + 2x^2 - y^3 &= 0, \\ 5xy - z^2 + 6 &= 0, \\ w^2 - w + 2x - y + z - 3 &= 0. \end{aligned}$$

For this system, there is no solution. To see this, note that the first equation implies that y is even. The second implies that z is even. The third equation implies a contradiction since $w^2 - w = w(w - 1)$ is always even.

These are special cases of the following problem: Is there an algorithm for deciding when a given polynomial equation with integer coefficients has a solution in the integers? This is Hilbert's tenth problem, to be discussed in Chapter 5. The answer, provided by the work of Martin Davis (born 1928), Yuri Matiyasevič (born 1947), Hilary Putnam (1926–2016), and Julia Robinson (1919–1985), is that there is no such algorithm. Our informal approach to computability is typically fine when describing an algorithm. However, as soon as we set out to show that an algorithm for completing some task cannot exist, a formal definition of an algorithm is required. To this end, we

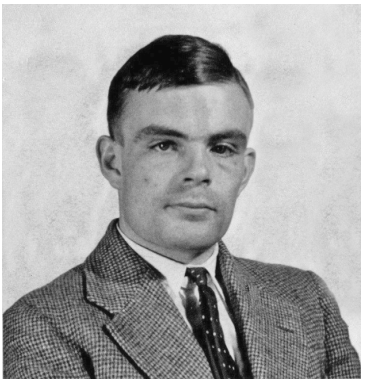
must inquire into what we mean by an algorithm. Our focus in this chapter will be on algorithms that can be used to compute a function.

Given the ancient origins of many algorithms, it is remarkable that a precise definition of a computable function was not given until the 1930s. As often happens in mathematics, numerous people were attempting to resolve this problem independently at around the same time. Gödel discussed *recursive functions* in 1934, and Alonzo Church (1903–1995) discussed the *lambda calculus* in 1936 (although soon after he too began to focus on recursive functions). In 1936 Alan Turing (1912–1954) discussed what are now called *Turing machines*. Eventually, it was shown that all three of these described the same class of functions, henceforth called the *computable functions*.

These definitions were proposed to capture the effectively computable functions. Church's assertion that the recursive functions did just that is called *Church's thesis*. Turing's assertion that the functions computable by Turing machines are the effectively computable functions is called *Turing's thesis*. Since these two theses are essentially one in the same, we will call it the *Church–Turing thesis*: a function is computable if and only it is effectively computable. Once we describe Turing machines, it will be clear that computable functions are effectively computable. However, the converse is not a statement that can be proven, given the informal definition of the effectively computable functions. There is, however, much evidence that the two are equivalent. In the end, effectively computable procedures all involve writing and erasing (or disregarding) symbols in various places based on the given problem and on previously written symbols, which is essentially the activity that Turing machines are designed to capture. In this text, we will generally be assuming the Church–Turing thesis.[1]

Alan Turing had a short life of only 41 years. As part of his undergraduate thesis, at the age of 22, he independently discovered Lindeberg's condition for the sufficiency of the central limit theorem (the necessity being proved later by Feller), a result of great importance in probability and statistics. His solution of the halting problem, which

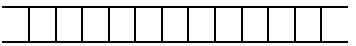
[1] In a slightly more advanced or specialized text, much can be done from the formal definition of computability alone, without having to appeal to the informal notion of *effectively computable functions*.



ALAN TURING (Photo courtesy of the University Archives, Princeton University Library.)

we discuss in this chapter, settled a question posed by Hilbert.[2] Turing's result came as a shock to Hilbert. During the Second World War, Turing was instrumental in breaking the Nazi Enigma code, which helped lead to an Allied victory. His innovation was to use statistical analysis which was a new technique in code-breaking. Sadly, he had a turbulent personal life and apparently committed suicide in 1954, although there is some debate as to the exact cause of his death [**Hod**]. Today, he is regarded as the founder of computer science, and much of the emerging theory of artificial intelligence depends on his ideas.

We now give a description of Turing machines. The machine consists of a movable device that can read and print symbols on an infinitely long strip of tape.[3] The tape, which extends infinitely in both directions, has been divided into a sequence of boxes:



[2]This question, posed by Hilbert in 1928, is called the *Entscheidungsproblem*. Essentially, it asked for an algorithm that would determine if a given statement is a theorem of a first order theory (which is discussed later in this chapter). Turing and Church independently answered this question in the negative in 1936. The negative solution to Hilbert's tenth problem, given in Chapter 5, provides another means to answer this question.

[3]Only a finite amount of tape will be used at any time, so one could instead imagine a finite strip of tape with the ability to increase the length of the tape at any time, if needed.

At any point in time, the machine is in one of finitely many *internal states*, denoted by $q_1, q_2, \ldots, q_m$. Each square on the tape contains one symbol from a previously specified *alphabet*: $s_0, s_1, \ldots, s_n$. We agree that the symbol s_0 will represent a blank square. To simplify the notation, we will write b for s_0 and 1 for s_1. Thus a blank tape is a tape with a b in each box. At any point, only finitely many boxes on the tape will not be blank. We let R and L denote a move by the machine to one box to the right and one box to the left, respectively. A *Turing machine* is a finite and nonempty set of quadruples of the form

$$q_i s_j s_k q_\ell, \quad q_i s_j R q_\ell, \quad \text{or} \quad q_i s_j L q_\ell.$$

No two quadruples in a Turing machine may contain the same first two symbols.

The quadruples in a Turing machine are instructions on what action the machine is to take when it is in internal state q_i and has just read symbol s_j. This is why we require no two quadruples to begin with the same first two symbols; otherwise, the machine would attempt to give two conflicting instructions in some situations. The quadruple $q_i s_j s_k q_\ell$ tells the machine to replace the symbol s_j with s_k and then enter internal state q_ℓ. The quadruples $q_i s_j R q_\ell$ and $q_i s_j L q_\ell$ tell the machine to move right and left, respectively, and then enter internal state q_ℓ.

Say our machine is acting on the tape

$\cdots$	s_{j_1}	s_{j_2}	s_{j_3}	s_{j_4}	s_{j_5}	$\cdots$

and is currently reading the symbol s_{j_3}. If the current state of the machine is q_i, we can denote this situation by

$$\cdots s_{j_1} s_{j_2} q_i s_{j_3} s_{j_4} s_{j_5} \cdots,$$

which is called an *instantaneous description* of the machine. The current state is written immediately preceding the symbol currently being read. We agree that an instantaneous description shall never end with a q_i, and we can always avoid this by appending an extra b to the end of the description if needed. If after an application of one instruction in the Turing machine, we move from instantaneous description α to instantaneous description β, we denote this by $\alpha \to \beta$.

For example, consider a Turing machine that consists of the two instructions $q_1 1 b q_2$ and $q_2 b L q_1$. Say the nonblank section of the tape consists of 111, so that the tape reads $\cdots b111b\cdots$, where the ellipses denote an infinite string of b's. We can always add more b's to the beginning and end of the instantaneous description, as required. Furthermore, suppose the machine is reading the second 1. Then our instantaneous description is $1q_1 11$. Since we are in state q_1 and have just read the symbol 1, we must replace the 1 with a b and move into state q_2. Thus our new instantaneous description is $1q_2 b1$. Continuing in this manner, we have

$$\begin{aligned} 1q_1 11 &\to 1q_2 b1 \\ &\to q_1 1b1 \\ &\to q_2 bb1 \\ &\to q_1 bbb1 \end{aligned}$$

(to get to the last line from the line above, remember that there are infinitely many b's before the first nonblank symbol on the tape). This machine stops, or *halts*, at this point as there is no instruction that begins $q_1 b$. When given the tape $\cdots b111b\cdots$ and started at the second 1, the machine halted on the tape $\cdots bbb1b\cdots$. We can see that this machine replaces a string of consecutive 1's with b's, starting at the initial position of the machine and then moving left, stopping when it runs out of consecutive 1's to replace.

Consider now a Turing machine that consists of the two instructions $q_1 1 b q_1$ and $q_1 b L q_1$. How does this machine differ from the machine above? Let's see.

$$\begin{aligned} 1q_1 11 &\to 1q_1 b1 \\ &\to q_1 1b1 \\ &\to q_1 bb1 \\ &\to q_1 bbb1 \\ &\to q_1 bbbb1 \\ &\to \cdots \end{aligned}$$

This machine also replaces 1's with b's (starting at the initial position of the machine) but this time does not halt. It will remove all 1's to

the left of the starting position, even those appearing after some blank boxes. Once it has replaced the leftmost 1 with a b, it will continue to move left forever.

Thus, we see that there are two possible outcomes when a Turing machine is applied to a tape. It may halt, or it may run forever, never halting. If the machine halts, then the resulting tape will be considered the output of the machine.

We would like to use Turing machines to compute functions defined on the natural numbers, and thus we need a way to represent these numbers on the tape. We use a block of $n+1$ consecutive 1's to represent the number n (we need to use one mark for 0 so that the machine can recognize that something is there!). We represent an n-tuple of natural numbers by inserting a blank between each number. For example, the 3-tuple $(3, 0, 1)$ is represented by $1111b1b11$. Given a Turing machine and an n-tuple $(m_1, \ldots, m_n)$, we agree to start the machine in state q_1 with initial position immediately preceding the first 1. If the machine halts, we can count the number of 1's on the resulting tape and call this the *result* of the computation. In this way we can define a *partial function* ϕ from $\mathbb{N}^n$ to $\mathbb{N}$: if the machine halts, $\phi(m_1, \ldots, m_n)$ is the result of the computation; otherwise, we leave $\phi(m_1, \ldots, m_n)$ undefined. ϕ is called a partial function because it may not be defined for all elements of $\mathbb{N}^n$.

Given a function f mapping from $\mathbb{N}^n$ to $\mathbb{N}$, we say f is *computable* if there is a Turing machine with associated function ϕ defined for all n-tuples (that is, ϕ is *total*) and $\phi = f$. If f maps from a subset of $\mathbb{N}^n$ to $\mathbb{N}$ and there is a Turing machine with associated partial function ϕ with $\phi = f$, we say f is *partially computable*.

As an example, we construct a Turing machine for the successor function $f(n) = n+1$. Since the number n is represented on the tape with $n+1$ consecutive 1's, we really want this machine to do nothing. However, there must be at least one instruction, as a Turing machine is nonempty. We can take the machine to consist of the single instruction q_1bbq_1. Then the machine begins with instantaneous description $q_1 1 \cdots 1$. Since there is no instruction that begins with $q_1 1$, the machine halts, leaving $n+1$ 1's on the tape, and so $\phi(n) = n+1$, as required.

We now construct a Turing machine for the addition function $A(m,n) = m+n$. The 2-tuple (m,n) will be represented by two strings of $m+1$ and $n+1$ consecutive 1's with a b in between, and we want to be left with $m+n$ 1's on the tape, not necessarily consecutive. Thus, we must construct a machine that removes two 1's. However, it must remove a 1 from each of the two strings of 1's, since in the case when m or n is 0 there will not be two consecutive 1's in a string. Let our Turing machine consist of the following instructions:

$$q_1 1 b q_1,\ q_1 b R q_2,\ q_2 1 R q_2,\ q_2 b R q_3,\ q_3 1 b q_3.$$

Writing 1^k to represent k consecutive 1's, we have

$$\begin{aligned} q_1 1 1^m b 1 1^n &\to q_1 b 1^m b 1 1^n \\ &\to b q_2 1^m b 1 1^n \\ &\to \cdots \\ &\to b 1^m q_2 b 1 1^n \\ &\to b 1^m b q_3 1 1^n \\ &\to b 1^m b q_3 b 1^n, \end{aligned}$$

at which point the machine halts, leaving $m+n$ 1's on the tape. This shows that the addition function is computable.

We give one more example. We will construct a doubling Turing machine that accepts a string of consecutive 1's as input and prints a second string consisting of the same number of 1's, with a b in between the two strings. This will show that the function $f(n) = 2(n+1)$ is computable. Furthermore, the output is our representation of the 2-tuple (n,n), on which we could apply another Turing machine. In our construction, we will use a third symbol, which we will write as 2. It will serve as a place marker. Our machine will change a 1 to a 2, add this 1 to the second string, and then return to the 2 and move it to the next 1. The instructions $q_1 1 2 q_1$ and $q_1 2 R q_2$ will change the first 1 to a 2 and move one space to the right. Then $q_2 1 R q_2$ and $q_2 b R q_3$ will move through any additional 1's and past one b. The instructions $q_3 1 R q_3$ and $q_3 b 1 q_4$ will move past any consecutive 1's (initially there will not be any) and then print a 1. Then instructions $q_4 1 L q_4$, $q_4 b L q_5$, $q_5 1 L q_5$, and $q_5 2 1 q_6$ will return us to the 2 and change it to a 1. A final instruction $q_6 1 R q_1$ moves us one place to the right and returns

us to our original state. If there is an additional 1 here, the process begins again. If there are no more 1's to copy, then we will be in state q_1 reading a b. As there is no instruction for this situation, the machine halts. In summary, our machine consists of instructions

$$q_1 1 2 q_1,\ q_1 2 R q_2,\ q_2 1 R q_2,\ q_2 b R q_3,\ q_3 1 R q_3,\ q_3 b 1 q_4,$$
$$q_4 1 L q_4,\ q_4 b L q_5,\ q_5 1 L q_5,\ q_5 2 1 q_6,\ q_6 1 R q_1.$$

We leave it as an exercise to verify that it takes instantaneous description $q_1 1^{n+1}$ to $1^{n+1} q_1 b 1^{n+1}$. If we would like to next apply a Turing machine to this representation of (n, n) on the tape, then we should leave the machine reading the first 1 and in an unused state. Adding the instructions $q_1 b L q_7$, $q_7 1 L q_7$, and $q_7 b R q_8$ will do this. If we now wish to apply an additional Turing machine, say addition, we may do so by changing all q_1 for that machine to q_8, all q_2 to q_9, and so on. In general, this is how we may apply one Turing machine after another. Doing this to our doubling machine with the machine for addition shows that the function $g(n) = n + n = 2n$ is computable.

We have shown that the successor function, the addition function, and the multiplication by 2 function are all computable functions. Many other common functions may be shown to be computable in this manner, such as the identity function and the multiplication function, although actually writing down the Turing machines may be cumbersome, as they may require many instructions. The subtraction function is defined to be $m - n$ when $m \geq n$ and is left undefined when $m < n$. Thus this is a partial function. It can be shown to be partially computable. Similarly, the truncated subtraction function, defined by $\text{sub}(m, n) = m - n$ for $m \geq n$ and $\text{sub}(m, n) = 0$ otherwise, can be shown to be computable, as can the difference function $|m - n|$.

Is it possible to create a Turing machine that accepts Turing machines as input? Turing machines take natural numbers as their input. To this end, we will assign a natural number to each Turing machine. Since the order of the quadruples in the Turing machine does not have any significance, we arbitrarily order them. Let's call them $I_1, I_2, \ldots, I_n$. Each I_m is of the form $q_i s_j A q_\ell$ where A is one of R, L, or s_k, and i, j, k, and ℓ are natural numbers with $i, \ell \geq 1$. To

this (ordered) Turing machine we assign the natural number

$$\prod_{m=1}^{n} p_{4m-3}^{i} p_{4m-2}^{j} p_{4m-1}^{\lambda_m} p_{4m}^{\ell},$$

where p_r is the rth prime number, and

$$\lambda_m = \begin{cases} 0 & \text{if } A \text{ is } R, \\ 1 & \text{if } A \text{ is } L, \text{ and} \\ k+2 & \text{if } A \text{ is } s_k. \end{cases}$$

Our first example of a Turing machine consisted of instructions $q_1 s_1 s_0 q_2$ and $q_2 s_0 L q_1$ (recall that we are writing b for s_0 and 1 for s_1). If we take the quadruples in the above written order, this machine is associated with the number $2^1 3^1 5^2 7^2 11^2 13^0 17^1 19^1$. If we take the instructions in the opposite order, the associated number is instead $2^2 3^0 5^1 7^1 11^1 13^1 17^2 19^2$. Thus this process does not assign a unique number to a Turing machine. However, once we assign an order to the quadruples (for example, lexicographical ordering), the number assigned is unique. Note that since no two quadruples in a Turing machine may contain the same first two symbols, not every natural number represents a Turing machine. For example, $49742 = 2^1 3^0 5^0 7^1 11^1 13^0 17^1 19^1$ yields quadruples $q_1 s_0 R q_1$ and $q_1 s_0 L q_1$, which is not a valid Turing machine. Furthermore, exponents on primes of the form p_{4m-3} and p_{4m} for $m \leq n$ must be at least 1. For example, $35 = 2^0 3^0 5^1 7^1$ would yield quadruple $q_0 s_0 L q_1$, which is not a valid Turing machine since q_0 is not a valid internal state. Given a natural number, we may find its prime factorization, write down the associated Turing machine quadruples, and determine if it is a valid Turing machine by checking that no q_0 appears and no two quadruples begin with the same two symbols. Using this, we may now create a Turing machine that acts on Turing machines. For example, one may construct a machine that, when given a natural number associated to a Turing machine, determines whether or not the machine contains an instruction that writes the symbol s_1. This is equivalent to determining if the given number is divisible by p_{4m-1}^3 for some m, but not any higher power of p_{4m-1}.

As we have seen, when a Turing machine is applied to some input, it may halt or may run forever. Is there a Turing machine that determines whether a given Turing machine halts on a given input? This is called the *halting problem.* In the same paper where he introduced Turing machines, Turing showed that no such machine exists.

Theorem 4.1 (Halting problem)**.** *There does not exist a Turing machine that can determine whether a given Turing machine halts on a given input.*

Proof. The first step is to enumerate the Turing machines. As we have already assigned a unique number to each Turing machine (assuming we list the instructions in lexicographical order), we may use this numbering to list the Turing machines, skipping any numbers that yield invalid Turing machines or previously listed machines (with the same instructions out of lexicographical order). Thus, we may list the Turing machines as T_1, T_2, and so on. We write $T_n(m)$ to designate the nth Turing machine applied to input m. Now suppose we were to possess a Turing machine $V(m, n)$ that determines whether $T_n(m)$ halts. We use this to construct a Turing machine U as follows. If V tells us that $T_n(n)$ does not halt, then $U(n)$ halts (on some output; it will not matter what it is). If V tells us that $T_n(n)$ halts, then $U(n)$ does not halt. This can be done by including a loop that runs forever. Thus, $U(n)$ halts if and only if $T_n(n)$ does not halt. We must have $U(n) = T_k(n)$ for some k. By the definition of U, $U(k)$ halts if and only if $T_k(k)$ does not halt. Since $U(k) = T_k(k)$, this is a contradiction. The Turing machine V cannot exist, and thus there is no Turing machine to determine if a given Turing machine halts on a given input. □

Note that the above theorem yields the nonexistence of a Turing machine that will correctly determine if *any* given Turing machine will halt on *any* given input. There may be Turing machines that can determine if Turing machines of a certain type will halt on certain inputs. However, as we have seen above, a single Turing machine that can determine the halting behaviour of all Turing machines on any input can be made to contradict itself. It is interesting to note

the similarities in the above argument with Cantor's diagonalization arguments from Chapter 1.

4.2. Recursive Functions

The Turing machines in the previous section can be thought of as an attempt to capture all functions that are calculated via mechanical means. In this section, we describe a class of functions that *ought to be* computable. This will be done inductively: we describe a few "basic" functions that are intuitively computable, and some "operations" used to create new functions from old that ought to pass on computability. This process can then be shown to yield the same computable functions of the previous section. This alternative description of the computable functions can be useful when proving results about computable functions; we shall be using it in Chapter 5 on Hilbert's tenth problem.

We define the *primitive recursive functions*. The *constant function* $C_0(x) = 0$, the *successor function* $S(x) = x + 1$, and the *projection functions* $P_i^n(x_1, \ldots, x_n) = x_i$ (for $n \geq 1$ and $1 \leq i \leq n$) are defined to be primitive recursive. Additional primitive recursive functions may be obtained by repeatedly applying to these functions the operations of composition and primitive recursion, which we will describe now. The *composition* of given function $f(t_1, \ldots, t_m)$ with given functions $g_1(x_1, \ldots, x_n), \ldots, g_m(x_1, \ldots, x_n)$ is the function

$$h(x_1, \ldots, x_n) = f(g_1(x_1, \ldots, x_n), \ldots, g_m(x_1, \ldots, x_n)).$$

Given functions $f(x_1, \ldots, x_n)$ and $g(t_1, \ldots, t_{n+2})$, *primitive recursion* yields the function $h(x_1, \ldots, x_n, z)$ satisfying

$$h(x_1, \ldots, x_n, 0) = f(x_1, \ldots, x_n) \quad \text{and}$$
$$h(x_1, \ldots, x_n, t+1) = g(t, h(x_1, \ldots, x_n, t), x_1, \ldots, x_n).$$

We may have $n = 0$ here, in which case we take f to be a constant. Then h is a function of one variable satisfying $h(0) = f$ and $h(t+1) = g(t, h(t))$.

These basic functions and function building operations yield a surprisingly large class of functions. For any given $k \in \mathbb{N}$, the function

$C_k(x) = k$ is primitive recursive, as it can be obtained from $C_0(x)$ by composing it with the successor function $S(x)$ k times. For example,

$$C_3(x) = S(S(S(C(x)))).$$

The addition function $A(x, y)$ can be obtained with primitive recursion on

$$f(x) = P_1^1(x) \quad \text{and} \quad g(t_1, t_2, t_3) = S(P_2^3(t_1, t_2, t_3)).$$

This yields $A(x, 0) = f(x) = x$ and

$$A(x, y + 1) = g(y, A(x, y), x) = S(A(x, y)) = (x + y) + 1.$$

We can obtain the multiplication function $M(x, y)$ with primitive recursion on

$$f(x) = C_0(x) \quad \text{and} \quad g(t_1, t_2, t_3) = A(P_2^3(t_1, t_2, t_3), P_3^3(t_1, t_2, t_3)).$$

This yields $M(x, 0) = C_0(x) = 0$ and

$$M(x, y + 1) = g(y, M(x, y), x) = M(x, y) + x.$$

When showing a function is primitive recursive, it is easiest to observe some sort of recursive behaviour and then find the functions f and g to fit this. For example, we know that $M(x, 0) = x \cdot 0 = 0$, which determines f, and

$$M(x, y + 1) = x(y + 1) = xy + x = M(x, y) + x,$$

which determines g. For example, let's show that the factorial function $F(x) = x!$ is primitive recursive. We want $F(0) = 1$, so we take $f = 1$ (since F is a function of one variable, we take $n = 0$ in the primitive recursion, and hence f is a constant). We also want

$$F(x + 1) = (x + 1)F(x) = M(S(x), F(x)).$$

This is to be equal to $g(x, F(x))$, and so we take

$$g(t_1, t_2) = M(S(t_1), t_2) = M(S(P_1^2(t_1, t_2)), P_2^2(t_1, t_2)).$$

Let $Q(x_1, \ldots, x_n)$ be a polynomial with natural number coefficients. Since Q is obtained with finitely many applications of addition

and multiplication on constants and the variables, it too is primitive recursive. For example, we have

$$\begin{aligned} 3x^2y^2 + 2 &= M(3, M(x^2, y^2)) + 2 \\ &= M(C_3(x), M(M(x,x), M(y,y))) + C_2(x) \\ &= A(M(C_3(x), M(M(x,x), M(y,y))), C_2(x)). \end{aligned}$$

The fact that polynomials with natural number coefficients are primitive recursive will be important to us in Chapter 5.

We would like to show that the truncated subtraction function, defined by $\mathrm{sub}(x,y) = x - y$ for $x \geq y$ and $\mathrm{sub}(x,y) = 0$ otherwise, is primitive recursive. To begin with, we look at the predecessor function defined by $\mathrm{pred}(x) = 0$ if $x = 0$ and $\mathrm{pred}(x) = x - 1$ if $x \geq 1$. This function is primitive recursive with $f = 0$ (f is a constant since pred is a function of one variable) and $g(t_1, t_2) = P_1^2(t_1, t_2)$. This yields $\mathrm{pred}(0) = 0$ and $\mathrm{pred}(x+1) = g(x, \mathrm{pred}(x)) = x$, as required. We may now show truncated subtraction is primitive recursive by taking

$$f(x) = P_1^1(x) \quad \text{and} \quad g(t_1, t_2, t_3) = \mathrm{pred}(P_2^3(t_1, t_2, t_3)).$$

This yields $\mathrm{sub}(x, 0) = P_1^1(x) = x$ and

$$\mathrm{sub}(x, y+1) = g(y, \mathrm{sub}(x,y), x) = \mathrm{pred}(\mathrm{sub}(x,y)),$$

as required. Note that since

$$|x - y| = A(\mathrm{sub}(x,y), \mathrm{sub}(y,x)),$$

it too is primitive recursive.

The function $Z(x)$ defined by $Z(0) = 1$ and $Z(x) = 0$ for $x \geq 1$ is primitive recursive, since $Z(x) = \mathrm{sub}(1, x)$. Now consider the primitive recursive function defined by

$$E(x,y) = Z(A(\mathrm{sub}(x,y), \mathrm{sub}(y,x))).$$

Using a result of the previous paragraph, this can be written as $Z(|x-y|)$. If $x \neq y$, then $E(x,y) = Z(|x-y|) = 0$ since $|x-y| \geq 1$. If $x = y$, then $E(x,y) = Z(|x-x|) = Z(0) = 1$. Thus, $E(x,y)$ is 1 if $x = y$ and is 0 if $x \neq y$. E is the *characteristic function* (or *indicator function*) for the equality relation. In this way we can extend the

definition of primitive recursion to relations: a relation is primitive recursive if its characteristic function is primitive recursive.

One may go on to show that many commonly encountered functions are primitive recursive, such as the greatest common divisor function and the nth prime function. However, showing this directly is beyond the scope of this text.

The goal of this section is to give an alternative description of the computable functions. While it can be shown that all primitive recursive functions are computable, it turns out that there are computable functions that are not primitive recursive. The simplest such example is the Ackermann function, given in 1928 by Wilhelm Ackermann (1896–1962), a student of Hilbert.[4] Let

$$\mathrm{Ack}(m,n) = \begin{cases} n+1 & \text{if } m = 0, \\ \mathrm{Ack}(m-1,1) & \text{if } m > 0 \text{ and } n = 0, \\ \mathrm{Ack}(m-1,\mathrm{Ack}(m,n-1)) & \text{if } m > 0 \text{ and } n > 0. \end{cases}$$

It is not difficult to use induction to show that

$$\begin{aligned} \mathrm{Ack}(1,n) &= n+2 = 2+(n+3)-3, \\ \mathrm{Ack}(2,n) &= 2(n+3)-3, \text{ and} \\ \mathrm{Ack}(3,n) &= 2^{n+3}-3. \end{aligned}$$

In each example above, the input is shifted by 3, an operation is performed, and then the output is shifted back by 3. In the first example, the operation is addition of 2. In the second, it is multiplication by 2. In the third, it is exponentiation with base 2. This pattern continues. Letting

$$\mathrm{EX}(n) = 2^{2^{\cdot^{\cdot^{\cdot^{2}}}}} \Big\} n \text{ 2's},$$

we have $\mathrm{Ack}(4,n) = \mathrm{EX}(n+3) - 3$ and $\mathrm{Ack}(5,n) = \mathrm{EX}_{n+3}(1) - 3$, where EX_n represents the function EX composed with itself n times.

[4] We actually give a variant of Ackermann's original function. Rózsa Péter (1905–1977) gave a modification of Ackermann's function in 1935, and in 1948 Raphael Robinson (1911–1995) further modified this function to the function we use here. Raphael Robinson was the husband of Julia Robinson, whose work was crucial to the solution of Hilbert's tenth problem.

The Ackermann function grows to be very large. For example,

$$\mathrm{Ack}(4,3) = 2^{2^{2^{2^2}}} - 3 = 2^{2^{65536}} - 3,$$

an extremely large number!

Although this function gets very large, it is computable, provided we ignore time and space constraints, as usual. To find $\mathrm{Ack}(m,n)$, we must compute $\mathrm{Ack}(m,n-1)$, $\mathrm{Ack}(m,n-2),\ldots,\mathrm{Ack}(m,0) = \mathrm{Ack}(m-1,1)$, along with numerous $\mathrm{Ack}(p,q)$ for $p<m$. Now we may be required to compute $\mathrm{Ack}(m-1,q)$ for some q quite large, but there are still only finitely many $\mathrm{Ack}(m-1,q)$ to compute. The computation of $\mathrm{Ack}(m,n)$ is of finite length. It turns out, however, that the Ackermann function grows too fast to be primitive recursive! It can be proved that for any primitive recursive function $g(x_1,\ldots,x_n)$, there is a natural number m such that $g(x_1,\ldots,x_n) < \mathrm{Ack}(m,\max\{x_i\})$ for all x_i.[5] If $\mathrm{Ack}(m,n)$ were primitive recursive, then there would exist some m_1 with $\mathrm{Ack}(m,n) < \mathrm{Ack}(m_1,\max\{m,n\})$ for all m,n. Taking $m = n = m_1$, we have $\mathrm{Ack}(m_1,m_1) < \mathrm{Ack}(m_1,m_1)$, a contradiction. Although many functions are primitive recursive, the collection of primitive recursive functions comes up short and is missing some computable functions.

Gödel, on a suggestion from Herbrand, added another operation to composition and primitive recursion. The functions produced with these operations are called the *general recursive* functions. Instead of using Gödel's formulation, we will discuss an equivalent formulation due to Stephen Cole Kleene (1909–1994), often called the *μ-recursive functions*. We will simply call these the *recursive* functions. Given a function $f(z,x_1,\ldots,x_n)$, our new operation *minimalization* yields the partial function

$$h(x_1,\ldots,x_n) = \min_z\{f(z,x_1,\ldots,x_n) = 0\}.$$

If no such z exists, then h is undefined. Adding minimalization to composition and primitive recursion and applying these to the constant function, successor function, and projection functions yields the *partial recursive functions*. If in addition we require the h produced

[5] An outline of the proof may be found in the exercises for Section 8.4 of [**Ho**].

by minimalization to be a total function, the resulting functions are called the *recursive functions*.

For example, let $h(x) = \min_z\{|x - (2z)| = 0\}$. If x is even, then $h(x) = x/2$, but if x is odd, then $h(x)$ is undefined. This is an example of minimalization yielding a partial function. This h is a partial recursive function but not a recursive function since it is not total.

We define functions $\mathrm{Quo}(a, b)$ and $\mathrm{Rem}(a, b)$ to be the quotient and remainder, respectively, when a is divided by $b + 1$ (we add 1 to b to avoid division by 0). Writing $a = q(b+1)+r$ with $0 \le r < b+1$, $\mathrm{Quo}(a, b) = q$ is the unique integer satisfying

$$q(b + 1) \le a < (q + 1)(b + 1).$$

Thus we can define $\mathrm{Quo}(a, b) = q$ to be the minimum q satisfying $a < (q + 1)(b + 1)$. Now, $Z(\mathrm{sub}(n, m))$ is equal to 0 when $m < n$ and is 1 otherwise, and hence

$$\begin{aligned}\mathrm{Quo}(a, b) &= \min_q\{Z(\mathrm{sub}((q + 1)(b + 1), a)) = 0\}\\ &= \min_q\{Z(\mathrm{sub}(M(S(q), S(b)), a)) = 0\}.\end{aligned}$$

Hence $\mathrm{Quo}(a, b)$ is recursive. Since $r = a - q(b + 1)$, it follows that

$$\mathrm{Rem}(a, b) = \mathrm{sub}(a, q(b + 1)) = \mathrm{sub}(a, M(\mathrm{Quo}(a, b), S(b)))$$

is also recursive.

It is possible to construct Turing machines to compute the constant function $C_0(x)$, successor function $S(x)$, and projection functions P_i^n. Furthermore, one may show that the operations of composition, primitive recursion, and minimalization preserve computability. That is, if $g_1, \ldots, g_m$, and f are computable by a Turing machine, then their composition $h = f(g_1, \ldots, g_m)$ is also computable by a Turing machine, and similarly for primitive recursion and minimalization. Hence if a function is recursive, then it is computable. Intuitively, this may not be too surprising. The computable functions are those for which we have means to calculate mechanically; the basic constant, successor, and projection functions are trivial to calculate, and it seems intuitively clear that our function building operations preserve mechanical calculation (remember that recursive functions

produced with minimalization are required to be total). What may be more surprising is that the converse to this statement also holds: if a function is computable, then it must be recursive. Unfortunately, a proof of this fact is outside of the scope of this book. These two results show that the recursive functions are one and the same as the computable functions. We previously saw that the Ackermann function is not primitive recursive. Since it is computable, it is a recursive function. Finally, we note that the partial recursive functions may be identified with the partially computable functions. A function produced via minimalization that is not total is akin to a Turing machine that does not halt on certain input.

We say a set S of natural numbers is *computably enumerable* if it is either empty or equal to the range of a computable function f. There are many other commonly used names for these sets, including *recursively enumerable*, *semidecidable*, and *listable*. We have an algorithm to list the elements of a computably enumerable set:

$$S = \{f(0), f(1), f(2), \ldots\}$$

for f a computable function. Conversely, if we possess an algorithm to list the elements of a set S, we can use it to give a computable function f with range S simply by setting the first listed element to be $f(0)$, the next to be $f(1)$, and so on.

If both S and its complement are computably enumerable, then we say that S is *decidable*. Other commonly used names for decidable sets include *recursive* and *computable*. Suppose S is decidable. Then given some natural number n, we have an algorithm to decide if n is a member of S or not: we alternately list the members of S and its complement. Since n is in one of these sets, it will appear after finitely many steps. Conversely, if we possess an algorithm to decide if an arbitrary natural number is a member of a set S, we can use it to list the members of S and its complement. Many commonly discussed sets of natural numbers are decidable. For example, the set of prime numbers is decidable: given an arbitrary natural number n, we may decide in finitely many steps whether or not n is prime (for example, we may use brute force trial division).

Clearly, every decidable set is also computably enumerable, but is the converse true? We have an algorithm to list the members of a computably enumerable set, but may not have an algorithm to determine whether or not an arbitrary natural number belongs to the set. Consider the following sequence of primes: $p_0 = 2$, and for $n > 0$, p_n is the least prime dividing $p_0 \cdots p_{n-1} + 1$. Since $p_0 \cdots p_{n-1} + 1$ has a remainder of 1 when divided by any p_i with $i \leq n - 1$, this sequence, called the *Euclid–Mullin sequence*, consists of distinct primes. Since we have an algorithm to list the members of the sequence, the set $S = \{p_n : n \in \mathbb{N}\}$ is computably enumerable. However, *we do not know* if S is decidable. There is no known algorithm that will determine if a given prime is a member of S. Say we wish to determine if 41 is a member of this set. All we can do right now is start listing the members of S and look for 41. If 41 is indeed a member of S, this process will end after finitely many steps. However, if 41 is not a member of S, then this process will never end. We do not currently have a way to decide if 41 will eventually appear on our list. To compute the members of the Euclid–Mullin sequence, we must multiply all previous elements, some of which are rather large. It is so computationally intensive that, at the time of this writing, only the first 51 elements of the sequence have been listed, and 41 is not among them. It is an open question as to whether S contains all prime numbers. Now, if someone were to prove this one day, then S would indeed be decidable. Thus S is an example of a computably enumerable set that may or may not be decidable. Is there a computably enumerable set that is provably not decidable? Our next theorem shows that such a set exists.

Theorem 4.2. *There exists a set of natural numbers that is computably enumerable but not decidable. Also, there exists a set of natural numbers that is not computably enumerable.*

Proof. In Chapter 1, we constructed a bijective function F from $\mathbb{N}$ to $\mathbb{N} \times \mathbb{N}$. One may now show that F is a computable function (in fact, this is Theorem 5.22 of Section 5.6). Applying F twice, we may traverse all ordered triples (m, n, x) of natural numbers. In particular, if we write $F(z) = (L(z), R(z))$, then

$$(L(z), L(R(z)), R(R(z)))$$

will range over all ordered triples of natural numbers as z ranges over the natural numbers. For each triple (m, n, x), run $T_n(m)$, the nth Turing machine on input m, for exactly x steps (if possible; it may halt at an earlier step). If the Turing machine has just halted on step x, then list (m, n). Thus the set

$$S = \{(m, n) \in \mathbb{N} \times \mathbb{N} : T_n(m) \text{ halts}\}$$

is computably enumerable. Suppose S were decidable. Then there is an algorithm, or Turing machine, that determines if $T_n(m)$ halts. By the halting problem (Theorem 4.1), this is a contradiction. Thus S is a computably enumerable set that is not decidable. It follows that

$$U = \{2^m 3^n : (m, n) \in S\}$$

is a set of natural numbers that is computably enumerable but not decidable.

Note that the complement of U is not computably enumerable. To see this, suppose the complement of U were computably enumerable. Then both U and its complement would be computably enumerable, and hence U would be decidable, which is not the case. □

4.3. Gödel's Completeness Theorems

In Chapters 1 and 2, we discussed Russell's paradox, which concerns the set $R = \{s : s \notin s\}$. The existence of such a set leads to a contradiction. This means we must set things up carefully in the hope that such a set cannot be constructed. We did this in Chapter 2, setting up set theory axiomatically with the ZFC axioms. If we have chosen our axioms correctly, it would be reasonable to hope that any statement of set theory may be proved or disproved with them. However, we have already mentioned that this is not the case. The continuum hypothesis cannot be proved or disproved using the axioms of ZFC set theory. How do we reach conclusions about what can and cannot be proved?

Another desired consequence of the correct choice of axioms is consistency. It would be reasonable to expect that if our axioms are chosen correctly, then we will never be able to prove a statement of the form "P and not P". For if we could prove such a statement, it

would follow that every statement is a theorem! To see this, let us prove a statement Q. Here Q can be any statement you please, provided it can be formalized in ZFC set theory, or whatever axiomatic system you are using. To prove Q, we use a proof by contradiction. Suppose Q does not hold. Then we have P and its negation, which is a contradiction. Thus Q holds. If every statement Q (and its negation!) were true, set theory would be trivial and useless. So, are the axioms of set theory consistent? Can we prove this?

In order to investigate these questions, we will separate the notions of proof (called a *syntactic* concept) and truth (called a *semantic* concept). It is natural to conflate these two notions, but some very interesting and important results can be deduced once we separate them and examine the interplay between them. We are especially interested in whether one of these notions implies the other, and when they overlap.

To begin, we discuss first order logic and first order theories. We will eventually work with two theories. One, called Peano arithmetic and abbreviated PA, is designed to capture arithmetic on the natural numbers. The other is ZFC set theory. First order theories are typically introduced with an interpretation in mind. For example, the intended interpretation of PA is arithmetic on the natural numbers. However, there may be other interpretations of the theory in addition to the intended interpretation.

In a first order theory, we can quantify over the variables, but not over sets of the variables. Since in ZFC the variables will be thought of as sets, this will present no real limitations. However, in PA the variables will be thought of as numbers. This will prevent us from quantifying over sets of numbers. Since, for example, functions and relations on the natural numbers are defined as sets, we are prevented from quantifying over all functions or over all relations, for example.

We would like to deduce some results *about* PA and ZFC, rather than merely prove theorems within them. Results of this type are called *metamathematical*, and they are part of a *metatheory*.

To begin, we discuss the language of a first order logic. A *first order language* consists of symbols, which we split into two types. The *logical symbols* are common to all first order languages. They

include variables x_1, x_2, and so on. The usual logical connectives of $\neg$ (negation), $\vee$ (disjunction), $\wedge$ (conjunction), $\Longrightarrow$ (implication), and $\Longleftrightarrow$ (the biconditional) are logical symbols, as are the universal quantifier $\forall$ and existential quantifier $\exists$.[6] The parentheses (and) are included for punctuation. Our final logical symbol is $=$ (equality). The second type of symbols in a first order language are called the *nonlogical symbols*. These may include constant symbols c_1, c_2, ..., n-ary function symbols F, and n-ary relation symbols R. The nonlogical symbols will differ between different first order languages. For example, in the language L_{ZFC} that we will use for ZFC set theory, we use just one 2-ary relation symbol: $\in$. We use no constant or function symbols in this language. In the language L_{PA} that we will use for PA, we use one constant symbol: 0. We use three function symbols and no relations. The successor function S is a 1-ary function, and $+$ and $\times$ are 2-ary functions.

We define the *terms* of our first order language inductively. Variables x_i and any constants are terms. If t_i are terms and F is an n-ary function symbol, then $F(t_1, \ldots, t_n)$ is a term. Thus in L_{ZFC}, our only terms are the variables x_i, which are to be thought of as sets. In L_{PA}, our terms include the variables, to be thought of as natural numbers. Also included are 0, $S(0)$, $S(S(0))$, and so on. Finally, if s and t are terms, then so are $s + t$, $s \times t$, and $S(s)$. For example, $S(S(x_1) + S(S(0)))$ is a term of L_{PA}.

Finally, we define the *formulas* of a first order language inductively. If s and t are terms, then $s = t$ is a formula. If t_i are terms and R is an n-ary relation, then $R(t_1, \ldots, t_n)$ is a formula. These first two types of formulas are called the *atomic formulas*. Finally, if P and Q are formulas, then $\neg P$, $P \vee Q$, $P \wedge Q$, $P \Longrightarrow Q$, $P \Longleftrightarrow Q$, $\forall x_n P$, and $\exists x_n P$ are formulas. Thus the atomic formulas of L_{ZFC} are of the form $x_i = x_j$ and $x_i \in x_j$ for variables x_i and x_j. The atomic formulas of L_{PA} are all of the form $s = t$ for terms s and t. If a formula has no free variables, it is called a *sentence*. We note that due to the way we have constructed the formulas of a language

[6] Note that a smaller set of logical connectives and quantifiers may be used, and the others may be defined in terms of these. For example, we might include only $\neg$, $\vee$ and $\forall$ as logical symbols. We may then define $P \wedge Q$ to mean $\neg(\neg P \vee \neg Q)$, $P \Longrightarrow Q$ to mean $\neg P \vee Q$, $P \Longleftrightarrow Q$ to mean $(P \Longrightarrow Q) \wedge (Q \Longrightarrow P)$, and $\exists x_n P$ to mean $\neg \forall x_n \neg P$.

L, there is an algorithm that will determine if a string of symbols of L is a formula.

We now give an overview of first order logic for a first order language L. This will allow us to discuss the syntactic notion of proof. We begin with a list of *logical axioms*. These may include formulas such as $\neg P \vee P$ (where P is a formula of L), $\forall x_1(x_1 = x_1)$, and $(\forall x_n P) \implies Q$ where Q is obtained from P by replacing all instances of x_n with a term t (provided P and t meet a technical condition[7]). Once we give a list of axioms, we describe our *rules of inference*, which will let us deduce new formulas from old. An example of a rule of inference is one that allows us to deduce $P \vee Q$ when we are given P. Once we have set these up, we can use the axioms and rules of inference to deduce a whole host of "derived" rules of inference, which are additional useful rules of inference that follow from the others. Different sources may use different axioms and rules of inference, but they end up with the same first order logic. As a rigorous treatment of first order logic is not the focus of this text, we do not attempt to list all axioms and rules of inference here. Suffice it to say, they include all the usual rules that you are used to using, such as modus ponens (from $P \implies Q$ and P, we may deduce Q), rules on substitution, rules on generalization (from P we may deduce $\forall x_n P$), and so on.

A proof of a formula P consists of a finite list of formulas, the last of which is P, and where each formula in the list is either an axiom or the result of applying a rule of inference on formulas that appear previously in the list. If there is a proof of P, we write $\vdash P$, and call P a *theorem*. For example, the following is a theorem of first order logic in any language that includes a 2-ary relation symbol R:

$$\neg\exists x\, \forall y\big(R(y,x) \iff \neg R(y,y)\big).$$

Note that if we take $R(x,y)$ to be $x \in y$ in L_{ZFC}, this is the negation of Russell's paradox. Now, since we have not explicitly listed the

[7]Namely, if y is a variable that appears in t, then no free occurrence of x_n in P can be quantified over with respect to y. Otherwise, this occurrence of y will become bound when it should not be. For example, take P to be $\exists y(x_n \in y)$ in L_{ZFC}. Intuitively, this says that x_n is an element of some set, which is the case for ZFC by the pairing axiom. Substituting $\varnothing$, say, for x_n yields $\exists y(\varnothing \in y)$, which does not present any problems. However, substituting y for x_n yields $\exists y(y \in y)$, which contradicts the axiom of regularity.

axioms and rules of inference of first order logic, we cannot write out a formal proof of the above sentence. However, we outline the steps here. First, we use the substitution axiom that we gave as an example of an axiom, which yields

$$\forall y\big(R(y,x) \iff \neg R(y,y)\big) \implies \big(R(x,x) \iff \neg R(x,x)\big).$$

Taking the contrapositive yields

$$\neg\big(R(x,x) \iff \neg R(x,x)\big) \implies \neg\forall y\big(R(y,x) \iff \neg R(y,y)\big).$$

Now, the premise in the above implication is always true. In particular, $\neg(P \iff \neg P)$ is a tautology. Thus we may use modus ponens to deduce

$$\neg\forall y\big(R(y,x) \iff \neg R(y,y)\big).$$

Using a generalization rule of inference, we may deduce

$$\forall x\neg\forall y\big(R(y,x) \iff \neg R(y,y)\big).$$

Finally, taking the double negation and using the fact that $\neg\forall x\neg P$ is the same as $\exists x P$ yields the result.

Now, we may wish to consider adding additional axioms to the usual axioms of first order logic. We will be doing this to obtain the theories of ZFC and PA. Let Γ be a set of sentences. A proof of P using Γ consists of a finite list of formulas, the last of which is P, and where each formula in the list is either an axiom of first order logic, an element of Γ, or the result of applying a rule of inference on formulas that appear previously in the list. If P has a proof using Γ, we write $\Gamma \vdash P$.

Note that so far we have only dealt with the syntactic side of first order logic. We have discussed proving theorems with first order logic but have not yet discussed truth. We now discuss the semantic side of first order logic. The following definition of truth is due to Alfred Tarski.

To begin, we define an interpretation I of a first order language. It consists of a nonempty set U_I. To each constant symbol c we associate an element c_I of U_I. To each function symbol F we associate a function F_I on U_I, and to each relation symbol R we associate a relation R_I on U_I. For example, for L_{PA} we may take U_I to be $\mathbb{N}$, 0_I to be the element 0 of $\mathbb{N}$, S_I to be the usual successor function on

$\mathbb{N}$, and $+_I$ and $\times_I$ to be the usual operations of addition and multiplication on $\mathbb{N}$. This is one possible interpretation of the language (and indeed it is the intended interpretation), but we are not limited to this interpretation.

Suppose we were to assign an element $\phi(x_i) \in U_I$ to each variable x_i. We may use this to assign truth values to the formulas. First, we may extend ϕ to the terms: $\phi(c) = c_I$ and $\phi(F(t_1, \ldots, t_n)) = F_I(\phi(t_1), \ldots, \phi(t_n))$ for terms t_i. We can now use ϕ to assign a truth value T or F (but not both) to each formula. In what follows, we describe when to assign the truth value T. Otherwise, we assign the truth value F. If s and t are terms, we set $\phi(s = t) = T$ if and only if $\phi(s)$ and $\phi(t)$ are the same elements of U_I. For terms t_i, we set $\phi(R(t_1, \ldots, t_n)) = T$ if and only if $(\phi(t_1), \ldots, \phi(t_n))$ is in the relation R_I. We set the truth values for formulas built out of logical connectives as one would expect:

$$\begin{aligned}
\phi(\neg P) = T &\text{ if and only if } \phi(P) = F.\\
\phi(P \vee Q) = T &\text{ if and only if } \phi(P) = T \text{ or } \phi(Q) = T.\\
\phi(P \wedge Q) = T &\text{ if and only if } \phi(P) = T \text{ and } \phi(Q) = T.\\
\phi(P \Longrightarrow Q) = T &\text{ if and only if } \phi(P) = F \text{ or } \phi(Q) = T.\\
\phi(P \Longleftrightarrow Q) = T &\text{ if and only if } \phi(P) = \phi(Q).
\end{aligned}$$

Finally, we set $\phi(\forall x_n P) = T$ if and only if $\psi(P) = T$ for any assignment ψ that may only differ from ϕ on x_n. That is, $\forall x_n P$ is true if P is true no matter how we assign x_n. We set $\phi(\exists x_n P) = T$ if and only if $\psi(P) = T$ for at least one assignment ψ that may only differ from ϕ on x_n.

In this way, assigning elements in U_I to the variables yields a truth value for each formula. A different assignment may change the truth values of some formulas. We say a formula P is *true in the interpretation* I if for every assignment ϕ in I we have $\phi(P) = T$, and we say P is *false in the interpretation* I if for every assignment ϕ in I we have $\phi(A) = F$. Note that it is possible that a formula with free variables may be neither true nor false in an interpretation. For example, take the formula $x_1 = x_2$, and say we use an interpretation with a set U_I that contains at least two elements. One possible assignment assigns all variables to the same element of U_I. In this

assignment, $x_1 = x_2$ is true. In a different assignment, x_1 and x_2 may be assigned to distinct elements of U_I, in which case $x_1 = x_2$ is false. Thus $x_1 = x_2$ is neither true nor false in our interpretation. On the other hand, if we used a single element set U_I in our interpretation, it would follow that $x_1 = x_2$ is true in this interpretation. Although it is possible that a formula may be neither true nor false in an interpretation, it cannot be both true and false. We also note that it can be proved that a sentence, which by definition has no free variables, is either true or false in an interpretation.

If Γ is a set of formulas, then we say an interpretation is a *model* of Γ if every formula of Γ is true in the interpretation. For example, although we have not yet given the axioms of PA, they will be chosen so that $\mathbb{N}$ is a model of PA.

So far, our notion of truth is tied to an interpretation. Perhaps a formula is true in some interpretations but not in others. If a formula P is true in *every* interpretation of the first order language, then we say that it is *logically valid*. As a simple example, the formula $x_1 = x_1$ is logically valid, as is the sentence $\forall x_1(x_1 = x_1)$. A less obvious example is that the following is logically valid:

$$\neg\exists x\,\forall y\big(R(y,x) \iff \neg R(y,y)\big).$$

This is the negation of Russell's paradox, which we previously saw is a theorem of first order logic. To see that it is logically valid, suppose χ is an assignment that when applied to the above formula yields F. Then

$$\chi\Big(\exists x\,\forall y\big(R(y,x) \iff \neg R(y,y)\big)\Big) = T.$$

Thus there is an assignment ϕ (that may only differ from χ on x) with

$$\phi\Big(\forall y\big(R(y,x) \iff \neg R(y,y)\big)\Big) = T.$$

Hence, we have

$$\psi\big(R(y,x) \iff \neg R(y,y)\big) = T,$$

where ψ is any truth assignment that can differ from ϕ only on y. In particular, we have $\psi(x) = \phi(x)$. Since we are free to choose the value of $\psi(y)$, let us take $\psi(y) = \phi(x)$ so that $\psi(x) = \psi(y)$. Now, by

our definition of assignments on logical connectives, we have

$$\psi(R(y,x)) \neq \psi(R(y,y)).$$

We break into two cases. Suppose $\psi(R(y,x))=T$ and $\psi(R(y,y))=F$. The latter implies that the relation R_I does not contain $(\psi(y),\psi(y))$, but the former implies that R_I contains $(\psi(y),\psi(x)) = (\psi(y),\psi(y))$. This is a contradiction. On the other hand, suppose $\psi(R(y,x)) = F$ and $\psi(R(y,y)) = T$. The former implies that the relation R_I does not contain $(\psi(y),\psi(x)) = (\psi(y),\psi(y))$, while the latter implies that R_I contains $(\psi(y),\psi(y))$, another contradiction. Thus the negation of Russell's paradox is both a theorem of first order logic and is logically valid.

We are now ready to explore the connections between the syntactic and semantic sides of theories of first order logic. Our first result is that we can prove only logically valid formulas. That is, a theorem of first order logic is true in every interpretation of the language.

Theorem 4.3 (Soundness theorem)**.** *If* $\Gamma \vdash P$*, then* P *is true in every model of* Γ*.*

Outline of proof. Recall that a model of Γ is an interpretation of the first order language in which every formula of Γ is true. The first step is to show that all of our logical axioms are logically valid. The next step is to show that each rule of inference preserves truth. That is, if the premises of the rule of inference are true in an interpretation, then so is the conclusion. Since we have not explicitly listed the axioms and rules of inference of first order logic in this text, we must skip the details for these first two steps. Now suppose we have a proof of P using Γ. This consists of a finite list of formulas P_i where each P_i is an axiom, an element of Γ, or the result of applying a rule of inference on previously listed formulas. Consider an arbitrary model of Γ. We give an inductive argument to show that each P_k is true in the model provided the P_i with $i < k$ are true in the model. If P_k is an axiom, then it is logically valid and, hence, true in the model. If P_k is an element of Γ, then it is true in the model by definition. If P_k is the result of applying a rule of inference to formulas P_i with $i < k$ and each P_i true in the model, then since the rules of inference preserve truth, P_k is true in the model. Thus each line of the proof

is true in the model. Since P is the last line of the proof of P, P is true in the model. □

Note that a model of the empty set is an interpretation of the first order language. Thus taking $\Gamma = \varnothing$ in the soundness theorem yields the following corollary.

Theorem 4.4. *Every theorem of first order logic is logically valid.*

We can use the soundness theorem to deduce results about consistency. Let Γ be a set of formulas. We say Γ is *consistent* if there is no formula P such that $\Gamma \vdash P$ and $\Gamma \vdash \neg P$.

Theorem 4.5 (Consistency theorem)**.** *If Γ has a model, then Γ is consistent.*

Proof. Suppose Γ has a model, but there is a formula P with $\Gamma \vdash P$ and $\Gamma \vdash \neg P$. The soundness theorem implies that P and $\neg P$ are true in the model. That is, P is both true and false in the model, a contradiction. □

Again, taking $\Gamma = \varnothing$ yields an important corollary.

Theorem 4.6. *First order logic is consistent. That is, there is no formula P such that both P and $\neg P$ are theorems of first order logic.*

We emphasize that for PA and ZFC, we will be adding to first order logic a list of additional axioms, which will make up the set Γ. Thus it is Theorem 4.5 that will apply to PA and ZFC; they are consistent if they have a model. We have not yet given the axioms of PA, but they will be chosen so that $\mathbb{N}$ is a model. Thus it follows that PA is consistent. However, some care must be taken here, as this is really giving us a result about *relative consistency*. Where can this model of PA be constructed? We have defined $\mathbb{N}$ in ZFC set theory. Thus, it is a theorem of ZFC that PA is consistent since PA has a model in ZFC. Now, ZFC has methods of proof that will be unavailable to us in PA. Do we need something as strong as ZFC to prove that PA is consistent? Is it possible that we can give a proof in PA of the consistency of PA? That is, can PA prove its own consistency? We will return to these questions in the next section.

We now discuss the converses to the consistency and soundness theorems, which were originally proved by Gödel. His original proof for the converse to the consistency theorem was greatly simplified by Leon Henkin (1921–2006). It is this method that we outline below.

Theorem 4.7 (Model existence theorem)**.** *If* Γ *is consistent, then* Γ *has a model.*

A rigorous proof of this theorem is outside of the scope of this text. Instead, we outline the main ideas. We wish to find a model of a consistent set Γ of formulas of a first order language L. An interpretation of Γ called the *canonical interpretation* is constructed, essentially out of the symbols of L, provided L has at least one constant symbol. However, it turns out that this interpretation is not necessarily a model of Γ. We say Γ is *complete* if for each sentence P, $\Gamma \vdash P$ or $\Gamma \vdash \neg P$. Henkin proved that if, in addition to being consistent, Γ is also complete and satisfies a technical condition called the *Henkin property*,[8] then the canonical interpretation of Γ is a model of Γ. Henkin then showed how to start with a consistent set Γ, add constants to L, and use these to extend Γ to a set Γ_1 that is consistent, complete, and Henkin. Then the canonical interpretation is a model of Γ_1. Discarding the newly added constant symbols leaves us with a model of Γ, completing the proof of the theorem.

We may now use the model existence theorem to prove the converse of the soundness theorem. We call this the adequacy theorem because it shows that the axioms and rules of inference of first order logic are adequate to prove all logically valid formulas.

Theorem 4.8 (Adequacy theorem)**.** *If* P *is true in every model of* Γ*, then* $\Gamma \vdash P$*.*

Proof. We prove the result for P a sentence. The result also holds when P has free variables, but proving this is just outside the scope of this text. We also assume the following result without proof, called the *extension theorem*: if P is not a theorem of Γ, then $\Gamma \cup \{\neg P\}$ is consistent. Now, suppose P is a sentence that is true in every model

[8] Γ is Henkin if the following is satisfied for each formula P with precisely one free variable x: if $\Gamma \vdash \neg\forall x P$, then there is a term t that contains no variables for which $\Gamma \vdash \neg Q$, where Q is obtained from P by replacing each instance of x with t.

of Γ, but Γ does not prove P. By the extension theorem, $\Gamma \cup \{\neg P\}$ is consistent. The model existence theorem then implies that $\Gamma \cup \{\neg P\}$ has a model. In this model, $\neg P$ is true and so P is false. On the other hand, this model is also a model of Γ. By our assumption, P is true in this model. This is a contradiction. □

Taking $\Gamma = \varnothing$ yields the converse of Theorem 4.4.

Theorem 4.9. *Every logically valid formula is a theorem of first order logic.*

Combining the consistency theorem with the model existence theorem and the soundness theorem with the adequacy theorem yields the two results known as Gödel's completeness theorems.

Theorem 4.10 (Gödel's completeness theorems)**.** Γ *is consistent if and only if* Γ *has a model. Furthermore,* $\Gamma \vdash P$ *if and only if* P *is true in every model of* Γ*.*

Gödel proved the model existence theorem in 1929 as part of his PhD dissertation. The completeness theorems were in response to questions raised by Hilbert earlier in the decade. Hilbert asked if every logically valid formula of first order logic has a formal proof. Gödel's completeness theorems answer this question in the affirmative. First order logic is complete in the sense that it can prove all of its logically valid formulas.[9]

In Chapter 2, it was mentioned that the axiom of choice is independent of the other axioms of set theory. It was also mentioned that the continuum hypothesis is independent of ZFC. We may now shed a little more light on these results. Let ZF denote all the axioms of ZFC set theory except for the axiom of choice, and let AC denote the axiom of choice. Suppose ZF is consistent. Then by the model existence theorem, it has a model. In 1938, Gödel used this to describe

[9]Note that the meaning of the word "complete" here is different from the way we used the word in our discussion of the model existence theorem. There, we defined a set Γ of sentences to be complete if for every sentence P, $\Gamma \vdash P$ or $\Gamma \vdash \neg P$. First order logic is not complete in this sense since, for example, $\forall x \forall y(x = y)$ is a sentence for which neither it nor its negation is a theorem of first order logic. This is because neither it nor its negation are logically valid: there are interpretations where it is true and interpretations where it is false. First order logic is complete in the sense that every formula that ought to be a theorem (that is, the logically valid formulas) is a theorem.

the *constructible universe*, a model of ZF in which both the axiom of choice and continuum hypothesis hold. Let us consider the axiom of choice. Since we now have a model of ZFC, it is consistent by the consistency theorem. Now suppose ¬AC were a theorem of ZF. Since adding an axiom to ZF will not change the proof of this theorem, it follows that ¬AC would be a theorem of ZFC. Trivially, AC is a theorem of ZFC. This contradicts the consistency of ZFC. In this way, Gödel showed that if ZF is consistent, then ZFC is also consistent, and hence ¬AC is not a theorem of ZF. In 1963, Cohen devised a new method of model building, called *forcing*, that he used to build models of ZF in which the negation of the axiom of choice and the negation of the continuum hypothesis hold. Thus, if ZF is consistent, it will remain consistent if we add to it ¬AC, and hence AC is not a theorem of ZF. Putting this together with Gödel's result, the axiom of choice is independent of ZF set theory. A similar argument may be made for the continuum hypothesis.

In the remainder of this section, we give two important results on models.

Theorem 4.11 (Löwenheim–Skolem theorem)**.** *If* Γ *has a model, then there is a countable model of* Γ*.*

By a countable model, we mean that the set U_I used in the interpretation is countable. The Löwenheim–Skolem theorem can be proved by using the proof of the model existence theorem, a proof that we have only summarized.[10] Here is the basic idea. Suppose Γ has a model. By the consistency theorem, Γ is consistent. In the proof of the model existence theorem, a model is constructed for a consistent set Γ. By the method of construction for this model (on which we omitted the details), the resulting model is countable.

Suppose ZFC is consistent. Then, by the model existence theorem, it has a model. We can apply the Löwenheim–Skolem theorem to obtain a countable model of ZFC. This consists of a countable set U_I and a relation $\in_I$ on this set. The axioms of ZFC set theory hold

[10]The theorem was originally proved in 1915 by Leopold Löwenheim (1878–1957). His proof was simplified by Thoralf Skolem (1887–1963) in the early 1920s. As this predates Gödel's completeness theorems, it is possible to prove the Löwenheim–Skolem theorem without the model existence theorem.

for the elements of U_I, and so all theorems of ZFC hold for the elements of this set. This includes Cantor's diagonalization argument that there is an uncountable set S. It seems that we have an uncountable set and yet only countably many elements that could possibly be in this set! This is called *Skolem's paradox*, although it is not really a paradox as it can be resolved. What does it mean to say that S is uncountable in our model? It means that there does not exist a bijective function f between the natural numbers and S. Recall that a function is, by definition, a set. Thus S really is a countable set, and the fact that S is uncountable in the model merely means that the set f is not an element of U_I.

Theorem 4.12 (Compactness theorem)**.** *Suppose every finite subset of* Γ *has a model. Then* Γ *has a model.*

Proof. Suppose that every finite subset of Γ has a model but that Γ does not. By the model existence theorem, it follows that Γ is inconsistent. Thus there is a formula P with $\Gamma \vdash P$ and $\Gamma \vdash \neg P$. Now, a proof has only finitely many lines, and so only finitely many formulas of Γ may be used in these two proofs. Let Γ_1 consist of these sentences. Then Γ_1 is not consistent. By the consistency theorem, Γ_1 does not have a model. Since we assumed that every finite subset of Γ has a model, this is a contradiction. □

We previously mentioned that although a theory is given with an intended interpretation in mind, there may be additional interpretations of the theory. That is, a theory Γ may have models other than the intended model. For example, consider PA. We will give the axioms of PA in the next section, but we have already discussed its language. The intended model of PA is $\mathbb{N}$ with the usual successor function and the usual operations of $+$ and $\times$. However, we can use the compactness theorem to give a very different model of PA, an example of what is called a *nonstandard model*. We introduce a new constant symbol c. To the axioms of PA, we adjoin infinitely many additional axioms: $c > 0$, $c > S(0)$, $c > S(S(0))$, and so on. Let us denote the set containing both the axioms of PA and these additional axioms by Γ. We write $S^n(0)$ as an abbreviation for n successive applications of the successor function to 0. In any finite subset of Γ,

only finitely many of the additional axioms may appear. Thus there is a largest n for which the axiom $c > S^n(0)$ is in the subset. If we interpret c as $n + 1$, then $\mathbb{N}$ is a model of this subset of Γ. Thus any finite subset of Γ has a model. By the compactness theorem, it follows that Γ has a model. Since the axioms of PA are a subset of Γ, this model is also a model of PA. However, this model contains an element that is larger than every successor of 0! Thus it is clear that this model is not the intended model $\mathbb{N}$.

Kurt Gödel (Photo courtesy Notre Dame Archives.)

We end this section with some words on the life of Gödel. Gödel's remarkable contribution was to make a mathematical distinction between truth and provability. This is the essence of his completeness theorem contained in his doctoral thesis. Gödel's incompleteness theorem, proved two years later and discussed in the next section, shattered the idyllic and romantic dream of mathematicians and philosophers from time immemorial (culminating in the lofty undertaking by

Whitehead and Russell in their three volume tome, *Principia Mathematica*) that once we properly formulate the rules for symbolic logic and choose the correct axioms, all universal truths will unfold. Gödel is now regarded as the greatest logician of all time.

Sadly, genius came at a price. Apparently, he was a hypochondriac all of his life and this ultimately lead to his death. In 1940, fleeing Nazi rule, he moved to the Institute for Advanced Study in Princeton along with his wife Adele, who was a nightclub dancer when they met. He was good friends with Einstein and in 1947, he became an American citizen. Studying the American Constitution, he found a logical inconsistency that could allow it to become a dictatorship!

In 1977, his wife, who would take meticulous care of Gödel's dietary needs, fell ill and had to be hospitalized. Gödel suffered mental depression and died of starvation in January 1978. Regarding his legacy, Manin [**AMS**, p. 36] writes "Gödel made clear that it takes an infinity of new ideas to understand all about integers only. Hence we need a creative approach to creative thinking, not just a critical one." We refer the reader to [**Daw**] for a readable biography of Gödel.

4.4. Gödel's Incompleteness Theorems

We now give the axioms of Peano arithmetic. Recall that the language for PA contains a constant symbol 0, a 1-ary function symbol S, and two 2-ary function symbols $+$ and $\times$.

PA1. $\forall x \neg(S(x) = 0)$.

PA2. $\forall x \forall y\big((S(x) = S(y)) \Longrightarrow (x = y)\big)$.

PA3. $\Big(P(0) \wedge \forall x\big(P(x) \Longrightarrow P(S(x))\big)\Big) \Longrightarrow \forall x P(x)$.

PA4. $\forall x(x + 0 = x)$.

PA5. $\forall x \forall y\big(x + S(y) = S(x + y)\big)$.

PA6. $\forall x(x \times 0 = 0)$.

PA7. $\forall x \forall y\big(x \times S(y) = (x \times y) + x\big)$.

PA3 is an axiom schema, as we have one axiom for each formula $P(x)$ in the language for PA. Here, $P(0)$ is obtained from $P(x)$ by

replacing each instance of x with 0. Similarly, $P(S(x))$ is obtained from $P(x)$ by replacing each x with $S(x)$.[11]

These axioms are given in the hope of capturing arithmetic on $\mathbb{N}$. The first axiom states that 0 is not the successor of any element. The second axiom states that if two elements have the same successor, then they are equal. Note that the converse of PA2 is a theorem of first order logic, and so it holds as well. The third axiom is giving us mathematical induction. Note that since in first order logic we cannot quantify over sets of the variable, we are forced to include a separate axiom for each formula of L_{PA}. The fourth and fifth axioms give the basic properties of addition, while the sixth and seventh give the basic properties of multiplication. As we saw in the comments following the compactness theorem of the previous section, these axioms have nonstandard models. However, $\mathbb{N}$ with the usual successor function and the usual operations of $+$ and $\times$ is a model of these axioms, and formal proofs of the important theorems of number theory on $\mathbb{N}$ may be given in PA.

We give an example of a proof in PA of a simple theorem. We show that

$$\forall x(0 + x = x).$$

Note that commutativity of addition is not an axiom of PA, and so we cannot deduce this theorem immediately from PA4. Commutativity of addition is a theorem of PA, but it must be proved from the axioms, and the theorem we are proving here is a step toward this. To prove this theorem, we will make use of PA3. Showing that $0 + 0 = 0$ and

$$(0 + x = x) \Longrightarrow (0 + S(x) = S(x))$$

(and then using a generalization rule of inference to apply the universal quantifier on x) will allow us to use PA3 to deduce the theorem. To show $0 + 0 = 0$, we use PA4 with 0 substituted for x. We now assume $0 + x = x$. By a rule of inference on functions from first order logic, we can apply S to both sides to get $S(0 + x) = S(x)$. Using PA5 with 0 substituted for x and then x substituted for y yields $0 + S(x) = S(0 + x)$. By transitivity of equality (another derived rule

[11] One may even take $P(x)$ to be a formula that does not contain x as a free variable. However, then $P(x)$, $P(0)$, and $P(S(x))$ are identical, and thus PA3 does not say anything interesting in this case.

of inference of first order logic), it follows that $0 + S(x) = S(x)$, as required.

Let Γ be a set of sentences in the language of PA. If Γ contains the axioms of PA, then we call Γ an *extension* of PA. If P is a theorem of PA and Γ is an extension of PA, then P is also a theorem of Γ.

If there is an algorithm that can determine if a given formula is a member of Γ, then we say Γ is *axiomatized*. For example, PA is axiomatized.

In 1930, Gödel proved and announced his first incompleteness theorem, which was published in 1931.

Theorem 4.13 (Gödel's first incompleteness theorem)**.** *If Γ is a consistent*[12] *axiomatized extension of PA,*[13] *then there is a sentence G such that neither G nor $\neg G$ is a theorem of Γ.*

When there is a sentence such that Γ can prove neither it nor its negation, we say Γ is *incomplete*. Thus Gödel's second incompleteness theorem shows that any consistent axiomatized extension of PA is incomplete.

Suppose Γ is a consistent axiomatized set of formulas, each of which is true in $\mathbb{N}$, the intended model of PA. Then $\mathbb{N}$ is also a model of Γ. If a sentence is a theorem of Γ, then by the soundness theorem it is true in the model $\mathbb{N}$. Gödel's first incompleteness theorem yields a sentence G such that both G and $\neg G$ are not theorems of Γ. Now, sentences are either true or false in a model, and so one of G or $\neg G$ is true in $\mathbb{N}$. Thus Gödel's theorem yields a sentence that is true in $\mathbb{N}$ but is not a theorem of the axiomatized extension of PA! In other words, our extension of PA has failed to capture all of arithmetic. One could argue that this sentence should then be added to Γ as an additional axiom. However, the incompleteness theorem could then be applied again to obtain a new sentence that is true in $\mathbb{N}$ but is not

[12] Gödel originally proved his theorem assuming something called ω-consistency, which is a stronger condition than consistency. In 1936, J. Barkley Rosser (1907–1989) showed how to replace this stronger condition with consistency.

[13] Actually, Gödel's first incompleteness theorem can be proved for a theory weaker than PA. Essentially, we can remove PA3, the induction axiom, but must then add a new symbol $<$ to the language and include a few additional axioms related to it. This theory is weaker than PA in the sense that every theorem it can prove is a theorem of PA, but there are theorems of PA that cannot be proved in this theory.

a theorem of this enlarged consistent axiomatized extension. There is no way that we can extend PA so that it is able to prove all sentences that are true in $\mathbb{N}$ and have it remain consistent and axiomatized.

We note that if we drop the requirement of consistency, then the result becomes trivial. An inconsistent extension of PA will prove every sentence in the language of PA, and is thus complete (but also useless). What if we drop the requirement of the extension being axiomatized? There does exist a complete and consistent extension of PA that is not axiomatized, called *complete arithmetic*. We simply let Γ be the set of all sentences in L_{PA} that are true in $\mathbb{N}$ (with the usual successor function and the usual operations of $+$ and $\times$ on $\mathbb{N}$). Since the axioms of PA are true in $\mathbb{N}$, this is an extension of PA. Trivially, every true sentence in $\mathbb{N}$ is a theorem of Γ, and so Γ is complete. It is also trivial that $\mathbb{N}$ is a model of Γ, and so Γ is consistent.[14] The fact that Γ is not axiomatized follows from Gödel's first incompleteness theorem: if it were axiomatized, then since it is consistent, it must not be complete, a contradiction. While a proof using complete arithmetic is trivial,[15] the problem with using the theory lies in determining exactly which sentences are elements of Γ!

A rigorous proof of Gödel's theorem is outside the scope of this book, but we are able to give some of the details. Gödel described a way to assign a natural number to each formula in L_{PA}, and then to each proof of a theorem in a consistent axiomatized extension Γ of PA. This allowed him to embed statements about provability into Γ itself. He showed how to use this to write down a sentence in L_{PA} that essentially says "I am not a theorem of Γ". He then showed that neither this sentence nor its negation is a theorem of Γ. Informally, one can see that if either the sentence or its negation were a theorem, it would lead to a contradiction. This is similar to the *liar paradox*: a person says "I am lying". Are they lying or telling the truth? Either situation leads to a contradiction. The contradiction arises because any sentence must be either true or false, at least in first order logic. The reason Gödel's sentence does not lead to a paradox

[14] As previously mentioned, the statement "Γ is consistent" is a theorem of a theory in which $\mathbb{N}$ can be constructed, such as ZFC set theory.

[15] Since every true sentence in $\mathbb{N}$ is an element of Γ, a proof of a true sentence P consists of one line: P!

is because there are three options. The first two, that the sentence or its negation are theorems of Γ, are ruled out. Thus it must be that Γ is incomplete.

We begin by explaining the Gödel numbering. To each of symbol s of the language of PA, we assign a number $f(s)$. The logical connectives $\neg$, $\vee$, $\wedge$, $\Longleftarrow$, and $\Longleftrightarrow$ are assigned numbers 1, 3, 5, 7, and 9, respectively. We assign to $\forall$ and $\exists$ the numbers 11 and 13. The left and right parentheses are assigned numbers 15 and 17. The equality symbol $=$ is assigned the number 19. For the nonlogical symbols, we set $f(0) = 21$, $f(S) = 23$, $f(+) = 25$, and $f(\times) = 27$. Finally, for $n \geq 1$, we set $f(x_n) = 27 + 2n$. Then, if P is a formula made up of symbols $s_1 s_2 \cdots s_n$, we set

$$g(P) = p_1^{f(s_1)} p_2^{f(s_2)} \cdots p_n^{f(s_n)},$$

where p_i is the ith prime number. For example, we have

$$g(\forall x_1(x_1 = x_1)) = 2^{11}3^{29}5^{15}7^{29}11^{19}13^{21}17^{17}.$$

On the other hand, if we are given a number, then there is an algorithm to determine which formula it represents, if any. We need only to factor the number, check that the number is the product of consecutive primes starting with 2, with each prime raised to an odd power, and, if so, then write down the symbols and apply the formula-recognizing algorithm. $g(P)$ is called the *Gödel number* for P. Note that for any formula P, $g(P)$ is even, although not every even number is the Gödel number of a formula.[16] Also note that the exponents on each prime in the factorization of a Gödel number are odd. Due to unique factorization, different formulas have different Gödel numbers; that is, g is injective.

Now suppose we have a proof using Γ of a formula P. The proof consists of a finite list of formulas $P_1, P_2, \ldots, P_n = P$. To this proof, we assign the Gödel number

$$p_1^{g(P_1)} p_2^{g(P_2)} \cdots p_n^{g(P_n)}.$$

Note that since each $g(P_i)$ is even, the Gödel numbers for formulas and for proofs are distinct.

[16] For example, $780 = 2^1 3^1 5^3$ is the Gödel number for $\neg\neg\vee$, which is nonsensical and not a formula. 10 is not a Gödel number since it is not the product of consecutive primes.

Gödel numbering allows us to give relations on the natural numbers that say something about proofs of theorems using Γ. For example, we can define a relation $R_{pr}(m, n)$ to hold if and only if m is the Gödel number of a formula P and n is the Gödel number of a proof of P using Γ. Since this is a relation on $\mathbb{N}$, can we find some way to represent this relation in L_{PA}?

For $m \in \mathbb{N}$, we write $\overline{m} = S^m(0)$, where $S^m(0)$ is an abbreviation for m successive applications of the successor function to 0. Thus $\overline{m}$ is a term in L_{PA}. In the model $\mathbb{N}$ of PA with the usual interpretation of the successor function, $+$, and $\times$, $\overline{m}$ is interpreted as the natural number m.

Given a relation $R(m_1, \ldots, m_n)$ on $\mathbb{N}$, we say that R is *expressible* in L_{PA} if there is a formula $P(x_1, \ldots, x_n)$ with free variables x_1, ..., x_n such that for all m_1, ..., m_n in $\mathbb{N}$, (1) if $R(m_1, \ldots, m_n)$, then $P(\overline{m}_1, \ldots, \overline{m}_n)$ is a theorem of PA, and (2) if $\neg R(m_1, \ldots, m_n)$, then $\neg P(\overline{m}_1, \ldots, \overline{m}_n)$ is a theorem of PA. We then say that P *expresses* R. One can give formal proofs in PA to show that many common relations on $\mathbb{N}$ are expressible in PA. For even fairly simple relations, these proofs can be long. Fortunately, Gödel proved that a large class of relations on $\mathbb{N}$ are expressible.

Given a relation $R(m_1, \ldots, m_n)$ on $\mathbb{N}$, we can define its *characteristic function* by

$$\mathrm{char}(m_1, \ldots, m_n) = \begin{cases} 1 & \text{if } R(m_1, \ldots, m_n), \\ 0 & \text{if } \neg R(m_1, \ldots, m_n). \end{cases}$$

We say that the relation R is *recursive* if its characteristic function is a recursive function. Gödel proved that if R is a recursive relation, then it is expressible. We are forced to omit the details here, but note that this takes some work to prove. Essentially, one must define what it means to express a function in PA, show that the basic recursive functions are expressible in PA, and then show that the operations of composition, primitive recursion, and minimalization preserve expressibility. Furthermore, these must be done with formal proofs in PA.

Let us return to the relation $R_{pr}(m, n)$ on $\mathbb{N}$. Suppose Γ is axiomatized. Given $m, n \in \mathbb{N}$, we can factor m to determine if it is the

Gödel number of a formula P in the language of PA. If it is, we can then factor n to determine if it is the Gödel number of a proof of P using Γ. This is where we use the fact that Γ is axiomatized. Thus we have an algorithm to determine whether or not $R_{pr}(m, n)$ holds, and hence to compute its characteristic function. By the Church–Turing thesis, the characteristic function, and hence the relation, is recursive.[17] By Gödel's theorem on expressibility, there is a formula $P_{pr}(x, y)$ of L_{PA} that expresses R_{pr}.

Gödel originally proved his first incompleteness theorem using something called ω-consistency instead of consistency. This condition is stronger than consistency. Rosser later showed how to weaken the ω-consistency condition to consistency. Γ is *ω-consistent* if for every formula $P(x)$ with precisely one free variable x, if $\Gamma \vdash P(\overline{n})$ for each $n \in \mathbb{N}$, then $\neg\forall x P(x)$ is not a theorem of Γ. To see that ω-consistency implies consistency, suppose Γ is ω-consistent and take $P(x)$ to be $x = x$. Since $\forall x(x = x)$ is an axiom of first order logic, $\overline{n} = \overline{n}$ is a theorem of Γ for each $n \in \mathbb{N}$. By ω-consistency, this implies $\neg\forall x(x = x)$ is not a theorem of Γ. If Γ were inconsistent, then every formula is a theorem of Γ. Thus Γ is consistent.

We are now ready to construct a sentence of PA such that neither it nor its negation is a theorem of an ω-consistent axiomatized extension Γ. We define a relation $R(m, n)$ on $\mathbb{N}$ to hold if and only if m is the Gödel number of a formula $P(x)$ that contains precisely one free variable x, and n is the Gödel number of a proof using Γ of $P(\overline{m})$. Note that we have introduced some self-reference, as $P(x)$ has Gödel number m, which we have placed back in P. Given $m, n \in \mathbb{N}$, we may factor m to determine if it is the Gödel number of a formula $P(x)$ with precisely one free variable. If it is, we may then factor n and use the fact that Γ is axiomatized to determine if n is the Gödel number for a proof using Γ of $P(\overline{m})$. Thus we have an algorithm to determine if R holds for m and n. By the Church–Turing thesis, this implies that R is recursive.[18] By Gödel's theorem on expressibility,

[17] One may show that R_{pr} is recursive without appealing to the Church–Turing thesis, provided that the condition that Γ is axiomatized is strengthened to Γ recursively axiomatized.

[18] The previous footnote about avoiding the Church–Turing thesis applies here as well.

there is a formula $Q(x, y)$ that expresses R. That is, for all $m, n \in \mathbb{N}$,

$$\text{if } R(m,n), \text{ then } \Gamma \vdash Q(\overline{m}, \overline{n}) \tag{4.1}$$

and

$$\text{if } \neg R(m,n), \text{ then } \Gamma \vdash \neg Q(\overline{m}, \overline{n}). \tag{4.2}$$

Let m be the Gödel number for the formula

$$\forall y \neg Q(x, y).$$

Note that this formula has precisely one free variable. Thus by the definition of R, for this value of m we have

$$\begin{gathered} R(m,n) \text{ if and only if } n \text{ is the Gödel number} \\ \text{for a proof using } \Gamma \text{ of } \forall y \neg Q(\overline{m}, y). \end{gathered} \tag{4.3}$$

Let G be the sentence

$$\forall y \neg Q(\overline{m}, y).$$

We will show that this is a sentence such that neither it nor its negation is a theorem of Γ. Note that G may be written as $\neg \exists y Q(\overline{m}, y)$. Thus it states that there is no $n \in \mathbb{N}$ such that $Q(\overline{m}, \overline{n})$. By (4.2), this means that there is no $n \in \mathbb{N}$ such that $R(m, n)$. By (4.3), there is no $n \in \mathbb{N}$ that is the Gödel number for a proof using Γ of $\forall y \neg Q(\overline{m}, y)$. However, this last sentence is simply G! Thus, informally, G states that G is not a theorem of Γ.

We now show that G is not a theorem of Γ. Suppose $\Gamma \vdash G$. Let n be the Gödel number of such a proof. By (4.3), $R(m, n)$ holds. By (4.1), $\Gamma \vdash Q(\overline{m}, \overline{n})$. On the other hand, since G is a theorem of Γ, we may take y to be $\overline{n}$, which shows that $\Gamma \vdash \neg Q(\overline{m}, \overline{n})$. This contradicts the consistency of Γ.

Finally, we show that $\neg G$ is not a theorem of Γ. Suppose $\Gamma \vdash \neg G$. Furthermore, suppose there were $n \in \mathbb{N}$ for which $R(m, n)$ holds. By the definition of R, n is the Gödel number of a proof using Γ of $\forall y \neg Q(\overline{m}, y)$. That is, $\Gamma \vdash G$, which contradicts consistency. Thus for all $n \in \mathbb{N}$, we must have $\neg R(m, n)$, and so $\Gamma \vdash \neg Q(\overline{m}, \overline{n})$. Since $\neg Q(\overline{m}, y)$ has precisely one free variable, we may use ω-consistency to deduce that $\neg \forall y \neg Q(\overline{m}, y)$ is not a theorem of Γ. That is, $\neg G$ is not a theorem of Γ, a contradiction.

As mentioned, in 1936 Rosser showed how to weaken the ω-consistency condition to consistency. This completes our discussion of Gödel's first incompleteness theorem, and we move on to his second incompleteness theorem.

Theorem 4.14 (Gödel's second incompleteness theorem)**.** *Let* Γ *be a consistent axiomatized extension of PA. Then the consistency of* Γ *is not a theorem of* Γ*.*

Essentially, the second incompleteness theorem says that any axiomatic theory strong enough to include Peano arithmetic cannot prove its own consistency. For example, PA itself cannot prove its own consistency. We previously saw that PA is consistent since $\mathbb{N}$ is a model. However, Gödel's second incompleteness theorem is telling us that this model cannot be constructed within PA. As we saw in Chapter 2, the set of natural numbers can be constructed in ZFC, and so there is a proof in ZFC of the consistency of PA. However, there can be no proof in PA of the consistency of PA.

While a rigorous proof of this theorem is far outside the scope of this book, we can provide some explanation on how it was proved. To start, how can we formulate the consistency of PA in L_{PA}? Recall that the relation $R_{pr}(m, n)$ holds if and only if m is the Gödel number of a formula P and n is the Gödel number of a proof of P using Γ. We observed that there is an algorithm to determine whether R_{pr} holds for any given m and n, which implies, by the Church–Turing thesis, that there is a formula $P_{pr}(x, y)$ of L_{PA} that expresses R_{pr}. Similarly, we define a relation $R_{for}(n)$ on $\mathbb{N}$ to hold if and only if n is the Gödel number of a formula of L_{PA}. Given n, we may factor it and determine if R_{for} holds for n. Thus, by the Church–Turing thesis, there is a formula $Q_{for}(x)$ of L_{PA} that expresses R_{for}.[19] Consider the following sentence of L_{PA}:

$$\exists x(Q_{for}(x) \wedge \neg \exists y P_{pr}(x, y)).$$

This sentence says that there is a formula that is not a theorem of Γ. Now, if Γ were inconsistent, then every formula would be a theorem

[19] As above, one may show that R_{for} is recursive without using the Church–Turing thesis if the conditions on Γ are strengthened.

of Γ. Thus the above sentence, which is in L_{PA}, is asserting that PA is consistent. Let us denote this sentence by C.

Let Γ be axiomatized. In the proof of his first incompleteness theorem, Gödel showed that if Γ is consistent, then the sentence G is not a theorem of Γ. Recall that, informally, G states that G is not a theorem of Γ. Gödel observed that the proof of his first incompleteness theorem can actually be carried out within Peano arithmetic![20] That is, Γ can prove that if Γ is consistent, then G is not a theorem of Γ, and so $\Gamma \vdash C \Longrightarrow G$. Suppose there is a proof using Γ of the consistency of Γ. That is, suppose $\Gamma \vdash C$. Applying modus ponens yields $\Gamma \vdash G$, which contradicts Gödel's first incompleteness theorem.

Gödel's second incompleteness theorem can be applied to any axiomatic system in which the axioms of PA are theorems. For example, it applies to ZFC set theory. If ZFC is consistent, then it cannot prove its own consistency.

As previously mentioned, since PA has a model that can be constructed in ZFC, the consistency of PA is a theorem of ZFC. Do we really need axioms as strong as those in ZFC to prove the consistency of PA? ZFC is stronger than PA in the sense that it can prove every theorem of PA. In order to prove the consistency of PA, do we need to move to a stronger theory? In 1936, Gerhard Gentzen (1909–1945) showed that the answer to both these questions is no. He gave a proof of the consistency of PA. His proof was carried out in a theory called *primitive recursive arithmetic*, augmented with something called *quantifier-free transfinite induction up to the ordinal* ϵ_0. This theory is much weaker than ZFC set theory. It is not stronger than PA, in the sense that there are theorems of PA that cannot be proved in this theory. It is also not weaker than PA, as it can prove the consistency of PA while PA cannot. A proof of the consistency of a axiomatized extension Γ of PA must be carried out in a *different*, but not necessarily stronger, theory.

[20] It requires a lot of work to actually demonstrate this. In fact, Gödel only asserted that this can be done and left it to others to carefully work out the details.

4.5. Goodstein's Theorem

Gödel's first incompleteness theorem yields a sentence of L_{PA} that is true in $\mathbb{N}$ but is not a theorem of PA (this is either G or $\neg G$ for G given in the previous section). Thus PA cannot prove all sentences that are true in $\mathbb{N}$. However, one could argue that this sentence is fairly "unnatural". It was constructed explicitly for the purpose of being true for $\mathbb{N}$ but unprovable, and one would hardly expect to arrive at such a sentence when doing "ordinary" mathematics with $\mathbb{N}$. In this section, we discuss a much more "natural" sentence that is true in $\mathbb{N}$ but cannot be proved in PA.

Given natural numbers m and n with $n \geq 2$, we may write m in the form

$$c_k n^k + c_{k-1} n^{k-1} + \cdots + c_1 n + c_0$$

for some natural number k, where $0 \leq c_i < n$ for each i. This is the base n expansion of m, and the c_i are the base n digits of m. For example, we have $521 = 2^9 + 2^3 + 1$ as the base 2 expansion of 521 (all missing powers of 2 have a coefficient of 0). Now, we can take the base n expansion of m and write the exponents in base n as well. We can do the same for any exponents appearing in these base n expansions, and so on. This process eventually terminates, leaving us with the *hereditary base n expansion of m*. For example, the hereditary base 2 expansion of 521 is

$$521 = 2^{2^{2+1}+1} + 2^{2+1} + 1.$$

Given a natural number $m > 0$, we define a sequence of natural numbers $G_n(m)$ for $n \geq 2$ called the *Goodstein sequence for m*. We set $G_2(m) = m$. To form $G_{n+1}(m)$, write $G_n(m)$ in its hereditary base n expansion, replace each n with $n + 1$, and then subtract 1. If a Goodstein sequence ever reaches 0, then it terminates.

The sequences $G_n(1)$, $G_n(2)$, and $G_n(3)$ quickly terminate. However, for larger m, this sequence seems to grow to be quite large. For

example,

$$\begin{aligned}
G_2(521) &= 521 = 2^{2^{2+1}+1} + 2^{2+1} + 1,\\
G_3(521) &= 3^{3^{3+1}+1} + 3^{3+1}\\
&\approx 10^{39},\\
G_4(521) &= 4^{4^{4+1}+1} + 4^{4+1} - 1\\
&= 4^{4^{4+1}+1} + 3\cdot 4^4 + 3\cdot 4^3 + 3\cdot 4^2 + 3\cdot 4 + 3\\
&\approx 10^{617},\\
G_5(521) &= 5^{5^{5+1}+1} + 3\cdot 5^5 + 3\cdot 5^3 + 3\cdot 5^2 + 3\cdot 5 + 2\\
&\approx 10^{10922}.
\end{aligned}$$

One can imagine how large $G_n(m)$ grows when starting with a large value of m.

In 1944, Reuben Goodstein (1912–1985) proved a surprising result about these sequences.

Theorem 4.15 (Goodstein's theorem)**.** *For any natural number $m > 1$, there is a natural number n such that $G_n(m) = 0$.*

The sequence $G_n(521)$ that seems to be growing so large will eventually reach 0. No matter what number we start with, the associated Goodstein sequence will eventually reach 0! We saw this happening for $m = 1$, 2, and 3. What about the next largest value of m, $m = 4$? We have $G_2(4) = 4 = 2^2$. Then

$$G_3(4) = 3^3 - 1 = 2\cdot 3^2 + 2\cdot 3 + 2.$$

In the next step, the 3's become 4's and 1 is subtracted. Following this, the 4's become 5's and another 1 is subtracted. This leaves us with

$$G_5(4) = 2\cdot 5^2 + 2\cdot 5.$$

In the next step, the 5's become 6's, which will yield a final term of $2\cdot 6$. We will have to take one of these two 6's to subtract 1, yielding

$$G_6(4) = 2\cdot 6^2 + 6 + 5.$$

Skipping ahead five steps brings us to

$$G_{11}(4) = 2\cdot 11^2 + 11$$

and

$$G_{12}(4) = 2 \cdot 12^2 + 12 - 1 = 2 \cdot 12^2 + 11.$$

Skipping ahead 11 steps brings us to

$$G_{23}(4) = 2 \cdot 23^2$$

and

$$G_{24}(4) = 2 \cdot 24^2 - 1 = 24^2 + 23 \cdot 24 + 23.$$

Repeating this process, we have

$$\begin{aligned} G_{47}(4) &= 47^2 + 23 \cdot 47, \\ G_{95}(4) &= 95^2 + 22 \cdot 95, \\ G_{191}(4) &= 191^2 + 21 \cdot 191. \end{aligned}$$

Note that when going from each of these values to the next, it takes one step to break into the final term so that we may subtract 1 (this reduces the coefficient of the final term by 1), and then in the subsequent steps we reduce the new final term, which is less than the base and hence does not grow, to 0. Thus, when going from each of the above given values to the next, the base is doubled and then increased by 1. Doing this 21 more times leaves us with just the square. At this point, the base is

$$2^{21} \cdot 191 + 2^{20} + 2^{19} + \cdots + 2 + 1 = 402653183.$$

Thus

$$\begin{aligned} G_{402653183}(4) &= 402653183^2, \\ G_{402653184}(4) &= 402653184^2 - 1 \\ &= 402653183 \cdot 402653184 + 402653183, \\ G_{805306367}(4) &= 402653183 \cdot 805306367. \end{aligned}$$

Repeating the above process, the coefficient 402653183 will decrease to 2 at base

$$\begin{aligned} B &= 2^{402653181} \cdot 805306367 + 2^{402653180} + 2^{402653179} + \cdots + 1 \\ &= 2^{402653181} \cdot 805306368 - 1. \end{aligned}$$

Then $G_B(4) = 2B$, and so

$$G_{B+1}(4) = 2(B+1) - 1 = (B+1) + B = 2B + 1,$$

where $2B + 1$ is the maximum value for this sequence. It stays here for a while,

$$G_{B+2}(4) = (B + 2) + B - 1 = 2B + 1,$$
$$G_{B+3}(4) = (B + 3) + B - 2 = 2B + 1,$$

and so on, until

$$G_{2B+1}(4) = 2B + 1,$$
$$G_{2B+2}(4) = 2B + 2 - 1 = 2B + 1.$$

Now our expansion contains only one coefficient, and it is less than the base. It will decrease by 1 after each step. Skipping ahead $2B$ steps, we have $G_{4B+2}(4) = 1$, and so $G_{4B+3}(4) = 0$. Thus $G_n(4)$ terminates for n equal to

$$\begin{aligned} 4B + 3 &= 4(2^{402653181} \cdot 805306368 - 1) + 3 \\ &= 2^{402653184} \cdot 402653184 - 1, \end{aligned}$$

reaching a maximum value of $2B+1 = 2^{402653182} \cdot 805306368 - 1$ along the way. The number of steps for $G_n(4)$ to terminate is approximately 7×10^{121210694}, an extremely large number. To put this in perspective, physicists today estimate the number of atoms in the universe to be about 10^{80}.

The above example hopefully gives some insight into what is happening with the Goodstein sequences. Even though the bases are increasing, subtracting 1 will chip away at the last coefficient, eventually requiring us to steal something from the preceding term.

How did Goodstein prove that every Goodstein sequence terminates? He used the ordinal numbers. Recall that in Section 2.2, we discussed the Cantor normal form: every ordinal may be written uniquely in the form

$$\omega^{\beta_1} c_1 + \omega^{\beta_2} c_2 + \cdots + \omega^{\beta_k} c_k,$$

where k and $c_1, \ldots, c_k$ are positive natural numbers, and $\beta_1 > \beta_2 > \cdots > \beta_k$ are ordinals. We may write each β_i in Cantor normal form, and so on. If the ordinal is less than ϵ_0, then this process will terminate, leaving us with what we will call the *hereditary Cantor normal form* of an ordinal. Recall also that if α and β are ordinals, then

$\alpha < \beta$ means $\alpha \in \beta$. Since the ordinals are well-ordered by $\in$, any strictly decreasing sequence of ordinals must be finite in length.

Fix $m > 1$. To each $G_n(m)$ we associate an ordinal as follows. Write $G_n(m)$ in hereditary base n expansion, with the coefficients multiplying powers of the base on the right. We then replace each occurrence of n with ω. For example, let us return to $G_n(521)$. We give the first four elements of the sequence along with their associated ordinals:

$$\begin{aligned}
G_2(521) &= 2^{2^{2+1}+1} + 2^{2+1} + 1 \rightsquigarrow \omega^{\omega^{\omega+1}+1} + \omega^{\omega+1} + 1,\\
G_3(521) &= 3^{3^{3+1}+1} + 3^{3+1} \rightsquigarrow \omega^{\omega^{\omega+1}+1} + \omega^{\omega+1},\\
G_4(521) &= 4^{4^{4+1}+1} + 3\cdot 4^4 + 3\cdot 4^3 + 3\cdot 4^2 + 3\cdot 4 + 3\\
&\qquad \rightsquigarrow \omega^{\omega^{\omega+1}+1} + \omega^{\omega}3 + \omega^3 3 + \omega^2 3 + \omega 3 + 3,\\
G_5(521) &= 5^{5^{5+1}+1} + 3\cdot 5^5 + 3\cdot 5^3 + 3\cdot 5^2 + 3\cdot 5 + 2\\
&\qquad \rightsquigarrow \omega^{\omega^{\omega+1}+1} + \omega^{\omega}3 + \omega^3 3 + \omega^2 3 + \omega 3 + 2.
\end{aligned}$$

This association yields ordinals less than ϵ_0 written in hereditary Cantor normal form. Furthermore, one can show that these ordinals are strictly decreasing (this is an exercise). Since any strictly decreasing sequence of ordinals is finite, the sequence of ordinals must terminate, and hence the sequence $G_n(m)$ must also terminate. However, the only way in which this can happen is if $G_n(m)$ reaches 0. Thus for any m, Goodstein's sequence must terminate at 0.

The above proof of Goodstein's theorem used infinite ordinal numbers to prove a result about a finite sequence of finite numbers. Goodstein's theorem may be formalized in Peano arithmetic. Since the theorem deals with a finite sequence of natural numbers, it seems reasonable to expect that one might be able to find a proof of it in PA. Amazingly, in the 1982 paper [**KP**] Laurence Kirby and Jeff Paris (born 1944) showed that this is not the case!

Theorem 4.16. *Goodstein's theorem is not a theorem of PA.*

The proof of this is far beyond the scope of this text. Essentially, they showed that Goodstein's theorem implies that PA is consistent.

Furthermore, they proved this entirely within PA. Thus, if Goodstein's theorem were a theorem of PA, we could apply modus ponens and deduce the consistency of PA as a theorem of PA. This contradicts Gödel's second incompleteness theorem, and so Goodstein's theorem cannot be a theorem of PA. Note that the negation of Goodstein's theorem is also not a theorem of PA. To see this, we note that since each axiom of PA is a theorem of ZFC, any proof carried out in PA can be translated into a proof of ZFC. Thus if PA could prove the negation of Goodstein's theorem, then so would ZFC. However, we have a proof of Goodstein's theorem in ZFC. The Kirby and Paris result on Goodstein's theorem was one of the first examples of a natural mathematical statement that is independent from PA.[21]

For $m > 0$, we define Goodstein's function $\mathcal{G}(m)$ to be the least value of n such that $G_n(m) = 0$. By Goodstein's theorem, this function is defined for all m. We have $\mathcal{G}(1) = 3$, $\mathcal{G}(2) = 5$, $\mathcal{G}(3) = 7$, and $\mathcal{G}(4) = 2^{402653184} \cdot 402653184 - 1$. This function grows very quickly. In a sense, it grows too quickly to be captured by PA. We previously discussed the Ackermann function, which also grows very fast. However, one can show that the Ackermann function can be captured by PA (in the sense that PA can prove that the Ackermann function is defined for all natural numbers, which PA cannot do for $\mathcal{G}$), and so Goodstein's function grows much faster than the Ackermann function.

Further Reading

There are many texts on computability and Turing machines. A classic is Davis's *Computability and Unsolvability* [**Da1**]. Although now 60 years old, it is still highly recommended. Our presentation of Turing machines was influenced by Davis's text.

Hodel's *An Introduction to Mathematical Logic* [**Ho**] is an excellent source for learning more on the material presented in Sections 4.3 and 4.4. Our treatment of these topics has certainly been influenced by Hodel's text. In fact, Hodel's text contains chapters on computability (using register machines instead of Turing machines),

[21] Paris and Leo Harrington (born 1946) gave a combinatorial example in 1977, a variant of Ramsey's theorem.

recursive functions, and Hilbert's tenth problems, making it an excellent source to learn more about the material covered in this text. However, note that these three topics are covered in later sections of the book, and may use the earlier material on first order logic.

A concise and well written survey of most of the topics contained in this chapter can be found in *What Is Mathematical Logic?* [**Cr**]. Each chapter is written by a different author: Crossley, Ash, Brickhill, Stillwell, and Williams.

There are numerous quality texts on mathematical logic. Enderton's *A Mathematical Introduction to Logic* [**En**] and Leary and Kristiansen's *A Friendly Introduction to Mathematical Logic* [**LK**] are both well regarded.

Exercises

4.1. Write a Turing machine that computes the constant function $C_0(n) = 0$.

4.2. Write a Turing machine that computes the identity function $f(m) = m$ on $\mathbb{N}$. Then write a machine for the identity function on $\mathbb{N}^n$.

4.3. Write a Turing machine that computes the function

$$Z(n) = \begin{cases} 1 & \text{if } n = 0, \\ 0 & \text{if } n \geq 1. \end{cases}$$

4.4. Write a Turing machine that computes $f(m, n) = |m - n|$.

4.5. Using the assignment of natural numbers to Turing machines described in the chapter, find the smallest natural number that is assigned to a valid Turing machine. Does this machine halt on any given nonempty input? Does it halt when started on an empty tape?

4.6. Show that there is no Turing machine that can determine if a given Turing machine acting on given input m will yield an output that contains the symbol s_k (for a fixed $k \geq 1$; since all but finitely many symbols on a tape are s_0, every Turing

machine will leave an s_0 on any input). That is, show that a machine $V(m, n)$ that determines if $T_n(m)$ yields an output that contains s_k leads to a contradiction. Note that if $T_n(m)$ does not halt, then $V(m, n)$ will answer in the negative (since there is no output, the output does not contain s_k). Model your argument on the solution of the halting problem. Compare this result to the Turing machine, mentioned in the chapter, that determines if a given Turing machine contains an instruction to write the symbol s_1.

4.7. Show that the function $P(x, y) = x^y$ is primitive recursive.

4.8. Given a primitive recursive function $p(x_1, \ldots, x_n, y)$, show that the function

$$h(x_1, \ldots, x_n, z) = \prod_{k=0}^{z} p(x_1, \ldots, x_n, k)$$

is primitive recursive.

4.9. Show that the "less than or equal to" relation on $\mathbb{N} \times \mathbb{N}$ is primitive recursive.

4.10. Use induction to show

$$A(1, n) = 2 + (n - 3) + 3,$$
$$A(2, n) = 2(n + 3) - 3, \text{ and}$$
$$A(3, n) = 2^{n+3} - 3.$$

Letting

$$\mathrm{EX}(n) = 2^{2^{\cdot^{\cdot^{\cdot^{2}}}}} \Big\} n \text{ 2s},$$

show $A(4, n) = \mathrm{EX}(n + 3) - 3$.

4.11. Show that if A is a decidable set, then so is its complement. Then show that if A and B are decidable sets, then so are $A \cup B$ and $A \cap B$.

4.12. Show that $(\forall x P \vee \forall x Q) \Longrightarrow \forall x (P \vee Q)$ (where P and Q are formulas that may or may not have x as a free variable) is logically valid. Is the converse logically valid?

4.13. Prove the converse of the compactness theorem.

4.14. Suppose P is true in every model of Γ. Prove that there is a finite subset Γ_1 of Γ such that P is true in every model of Γ_1.

4.15. Let P_2 be the sentence $\exists x_1 \exists x_2(\neg(x_1 = x_2))$, let P_2 be the sentence $\exists x_1 \exists x_2 \exists x_3(\neg(x_1 = x_2) \wedge \neg(x_1 = x_3) \wedge \neg(x_2 = x_3))$, and so on. Then if P_n is true in a model, that model must contain at least n elements. Suppose Γ has arbitrarily large finite models. Let

$$\Gamma_1 = \Gamma \cup \{P_n : n \in \mathbb{N}, n \geq 2\}.$$

Show that every finite subset of Γ_1 has a model. Deduce that Γ must have an infinite model.

4.16. Let Γ be axiomatized, and let $R_{pr}(m, n)$ be the relation on $\mathbb{N}$ that holds if and only if m is the Gödel number of a formula P of L_{PA} and n is the Gödel number of a proof of P using Γ. We saw that this is a decidable relation. Let $P_{pr}(x, y)$ be a formula of L_{PA} that expresses $R_{pf}(m, n)$. Let Q be an axiom of PA with Gödel number m. Is

$$P_{pr}(\overline{m}, \overline{2^m})$$

true or false in Γ?

4.17. Let $G_n(m)$ be the Goodstein sequence for m. Write out all elements of the sequences $G_n(1)$, $G_n(2)$, and $G_n(3)$.

4.18. Let

$$\alpha = \omega^{\gamma_1} a_1 + \omega^{\gamma_2} a_2 + \cdots + \omega^{\gamma_k} a_k$$

and

$$\beta = \omega^{\delta_1} b_1 + \omega^{\delta_2} b_2 + \cdots + \omega^{\delta_\ell} b_\ell$$

(with k and ℓ positive natural numbers, $a_i, b_i \in \mathbb{N}$, and γ_i, δ_i ordinals with $\gamma_1 > \gamma_2 > \cdots > \gamma_k$ and $\delta_1 > \delta_2 > \cdots > \delta_\ell$) be two ordinals in Cantor normal form, both less than ϵ_0. Describe the conditions on k, ℓ, γ_i, and δ_i that are equivalent to $\alpha < \beta$. Use this to show that the ordinals associated to the elements of a Goodstein sequence are strictly decreasing.

Chapter 5

Hilbert's Tenth Problem

5.1. Diophantine Sets and Functions

In the tenth of his previously mentioned list of twenty-three problems, Hilbert asked for an algorithm to determine if an arbitrary polynomial equation with integral coefficients and any number of variables (called a *Diophantine equation*) has integer solutions. Here is his wording, translated to English:

> Given a Diophantine equation with any number of unknown quantities and with integral numerical coefficients: to devise a process according to which it can be determined by a finite number of operations whether the equation is solvable in integers.

In this chapter, we will show that no such algorithm exists. Many of these results were given by Martin Davis, Julia Robinson, and Hilary Putnam in the 1950s and 1960s. The final missing piece was given in 1970 by Yuri Matiyasevič. In order to increase readability and ease of understanding, we do not present the solution to Hilbert's tenth problem in the order of discovery. Instead, we will comment on the history of the solution to the problem at the end of the chapter.

In what follows, we will discuss algorithms for determining if a Diophantine equation has natural number solutions. The problem

Martin Davis, Julia Robinson, and Yuri Matiyasevič (Photograph taken by Louise Guy in Calgary, 1982. Courtesy Richard Guy.)

is equivalent to that for integer solutions. In particular, to test the equation $P(x_1, \ldots, x_n) = 0$ for natural number solutions $(x_1, \ldots, x_n)$, we may use the fact from Theorem 3.15 that every natural number can be written as the sum of four squares and test the equation

$$P(p_1^2 + q_1^2 + r_1^2 + s_1^2, \ldots, p_n^2 + q_n^2 + r_n^2 + s_n^2) = 0$$

for integer solutions $(p_1, q_1, r_1, s_1, \ldots, p_n, q_n, r_n, s_n)$. Conversely, to test the equation $P(z_1, \ldots, z_n) = 0$ for integer solutions $(z_1, \ldots, z_n)$, we may test the equation $P(x_1 - y_1, \ldots, x_n - y_n)$ for natural number solutions $(x_1, y_1, \ldots, x_n, y_n)$.

When working with Diophantine equations, we typically have an equation in mind and would like to determine its solutions. Here, we turn the problem on its head: we start with a set of "solutions", and would like to find a corresponding Diophantine equation. We say a set S of ordered n-tuples of natural numbers is *Diophantine* if there is a natural number m and a polynomial $P(x_1, \ldots, x_n, y_1, \ldots, y_m)$ with integer coefficients such that $(x_1, \ldots, x_n)$ is in S if and only if there exist natural numbers $y_1, \ldots, y_m$ for which $P(x_1, \ldots, x_n, y_1, \ldots, y_m) = 0$. That is,

$$(x_1, \ldots, x_n) \in S \iff (\exists y_1, \ldots, y_m)(P(x_1, \ldots, x_n, y_1, \ldots, y_m) = 0).$$

Note that P may, and in all but trivial cases will, have negative coefficients, but the x_i and y_i are natural numbers.

We give some examples of Diophantine sets. The set of natural numbers is trivially Diophantine, as

$$x \in \mathbb{N} \iff (\exists y)(x - y = 0).$$

Any finite set is Diophantine, as

$$x \in \{s_1, \ldots, s_n\} \iff (x - s_1) \cdots (x - s_n) = 0.$$

The set of composite numbers is Diophantine, since x is composite if and only if there exist y and z with $x = (y+2)(z+2)$. Thus we may take

$$P(x, y, z) = x - (y + 2)(z + 2).$$

The set of numbers that are not powers of 2 is Diophantine. This is because a number is not a power of 2 if and only if it is divisible by an odd number greater than 1. That is, x is not a power of 2 if and only if there exist y and z with $x = y(2z+3)$, and hence we may take

$$P(x, y, z) = x - y(2z + 3).$$

Since relations are sets, they too may be Diophantine. For example, the $\leq$ relation is Diophantine since $x \leq y$ if and only if there exists z such that $x + z = y$. The divisibility relation is Diophantine since $x \mid y$ if and only if there exists z such that $xz = y$. What about the "congruence modulo 5" relation?[1] We know $x \equiv y \pmod 5$ if and only if $5 \mid (x-y)$. However, note that $x-y$ is not necessarily a natural number, as it could be negative. We can break into two cases: $x \equiv y \pmod 5$ if and only if there exists a natural number z with $x-y = 5z$ or $y - x = 5z$. That is, with $x - y - 5z = 0$ or $y - x - 5z = 0$. We can combine this into a single equation with multiplication, since the product of two numbers is 0 if and only if at least one of the numbers is 0. Thus $x \equiv y \pmod 5$ if and only if there exists z such that

$$(x - y - 5z)(y - x - 5z) = 0.$$

We now show that the set of numbers that are both composite and not a power of 2 is Diophantine. By our work above, x is composite if and only if there exist y_1 and z_1 such that $x - (y_1 + 2)(z_1 + 2) = 0$. Similarly, x is not a power of 2 if and only if there exist y_2 and z_2 such that $x - y_2(2z_2 + 3) = 0$. These two equations may be combined

[1] There is nothing special about 5 here; any other moduli may be used.

into a single equation by using the fact that the sum of the squares of two numbers is 0 if and only if both numbers are 0. Thus x is both composite and not a power of 2 if and only if there exist y_1, y_2, z_1, z_2 with

$$(x - (y_1 + 2)(z_1 + 2))^2 + (x - y_2(2z_2 + 3))^2 = 0.$$

As seen in the last two examples, given expressions that yield Diophantine sets, we may combine them using the logical connectives "and" (conjunction) and "or" (disjunction) to obtain another Diophantine set. Since both $5 \mid (x - y)$ and $5 \mid (y - x)$ may be given Diophantine definitions, so can their disjunction. Since we may give Diophantine definitions for the composite numbers and the numbers that are not powers of 2, we may also do so for their conjunction. This is because

$$(\exists y_1, \ldots, y_m)(P_1 = 0) \text{ and } (\exists z_1, \ldots, z_n)(P_2 = 0)$$
$$\Longleftrightarrow$$
$$(\exists y_1, \ldots, y_m, z_1, \ldots, z_n)(P_1^2 + P_2^2 = 0)$$

and

$$(\exists y_1, \ldots, y_m)(P_1 = 0) \text{ or } (\exists z_1, \ldots, z_n)(P_2 = 0)$$
$$\Longleftrightarrow$$
$$(\exists y_1, \ldots, y_m, z_1, \ldots, z_n)(P_1 P_2 = 0).$$

There is another logical symbol that may be freely used: the existential quantifier $\exists$. For example, the set S of ordered pairs (x, y) with $y^2 \mid x$ is Diophantine because

$$(x, y) \in S \iff (\exists z)(y^2 z - x = 0).$$

Since the expression $y^2 | x$ yields a Diophantine set, so does $(\exists y)(y^2 | x)$. To see this, let $T = \{x : \exists y(y^2 \mid x)\}$. We then have

$$x \in T \iff (\exists y, z)(y^2 z - x = 0).$$

The composite numbers and numbers that are not powers of 2 are Diophantine. What about the prime numbers or the powers of 2? Are these sets Diophantine? It turns out that they are but the proof is not easy, as we shall soon see.

Hilary Putnam found the following surprising characterization of Diophantine sets of natural numbers.

Theorem 5.1. *A set S of natural numbers is Diophantine if and only if there is a polynomial P such that S is equal to the set of natural numbers in the range of P. That is, if and only if S is the nonnegative range of P.*

Proof. First let S be the natural numbers in the range of a polynomial $P(x_1, \ldots, x_m)$. Then $x \in S$ if and only if $\exists x_1, \ldots, x_m$ such that $x = P(x_1, \ldots, x_m)$. That is, if and only if $P(x_1, \ldots, x_m) - x = 0$, so that S is Diophantine. Conversely, let S be Diophantine, so that

$$x \in S \iff (\exists y_1, \ldots, y_m)(Q(x, y_1, \ldots, y_m) = 0)$$

for some polynomial Q. Let

$$P(x, y_1, \ldots, y_m) = (x+1)(1 - Q^2(x, y_1, \ldots, y_m)) - 1.$$

We show that S is equal to the nonnegative range of P. If $x \in S$, choose $y_1, \ldots, y_m$ such that $Q(x, y_1, \ldots, y_m) = 0$. Then

$$P(x, y_1, \ldots, y_m) = x$$

and so x is in the range of P. On the other hand, if $z = P(x, y_1, \ldots, y_m)$ with $z \geq 0$, then

$$z + 1 = (x+1)(1 - Q^2(x, x_1, \ldots, x_m)).$$

Since $z + 1 \geq 1$ and $x + 1 \geq 1$, $Q(x, x_1, \ldots, x_m)$ must vanish. This implies $z = x$ and $x \in S$. □

One nice aspect of the above result is that it is constructive. If we know the polynomial in the Diophantine definition of a set, then we can construct the *nonnegative range* polynomial. Take the set of composite numbers. We saw that x is composite if and only if there exist y and z with $x - (y+2)(z+2) = 0$. Taking the left-hand side of this equation as our Q, we see that the set of composite numbers is the nonnegative range of

$$(x+1)\Big(1 - \big(x - (y+2)(z+2)\big)^2\Big) - 1.$$

When one stares at this equation long enough, it becomes clear why this is, although the clarity is lost if we are given the equation in expanded form:

$$\begin{gathered}-xy^2z^2 - y^2z^2 + 2x^2yz - 4xy^2z - 4xyz^2 - x^3 + 4x^2y + 4x^2z \\ -4xy^2 - 4y^2z - 4xz^2 - 4yz^2 - 14xyz + 7x^2 - 4y^2 - 4z^2 \\ -12xy - 12xz - 16yz - 7x - 16y - 16z - 16.\end{gathered}$$

It is not at all obvious that the set of nonnegative values taken on by the above polynomial is equal to the set of composite numbers![2] The consequences of this result are even more surprising once we show that the prime numbers are Diophantine and that the powers of 2 are Diophantine. It is then possible to give a polynomial for which the nonnegative range is precisely the set of prime numbers and a polynomial for which the nonnegative range is precisely the powers of 2!

A function f mapping from $\mathbb{N}^n$ to $\mathbb{N}$ is called *Diophantine* if the set

$$\{(x_1, \ldots, x_n, y) : y = f(x_1, \ldots, x_n)\}$$

is a Diophantine set. That is, f is Diophantine if its *graph* is Diophantine. One of our goals will be to determine which functions are Diophantine.

In Chapter 1 we discussed Cantor's pairing function, which is a bijection P from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{N}$ with formula

$$P(x, y) = \frac{(x + y)(x + y + 1)}{2} + x.$$

We let $F = P^{-1}$ and write $F(z) = (L(z), R(z))$ so that

$$P(L(z), R(z)) = z, \quad L(P(x, y)) = x, \quad \text{and } R(P(x, y)) = y.$$

Then $P(x, y)$, $L(z)$, and $R(z)$ are Diophantine functions since

$$\begin{aligned} z = P(x, y) &\iff 2z = (x + y)(x + y + 1) + 2x, \\ x = L(z) &\iff (\exists y)(2z = (x + y)(x + y + 1) + 2x), \text{ and} \\ y = R(z) &\iff (\exists x)(2z = (x + y)(x + y + 1) + 2x). \end{aligned}$$

[2]Note that this polynomial may take on all sorts of negative values, including negative prime numbers. However, the set of all nonnegative numbers in its range is precisely the set of composite numbers.

We also note that $L(z) \leq z$ and $R(z) \leq z$.

In Chapter 1, we gave the chart below. If the row number is taken to be x and the column number to be y, then the body gives $P(x, y)$. We can now use this chart to find $L(z)$ and $R(z)$. We find z in the body of the chart. Then $L(z)$ is the row number for this entry, and $R(z)$ is the column number. For example, since 22 is in the row for 1 and the column for 5, we have $L(22) = 1$ and $R(22) = 5$.

		$R(z)$					
		0	1	2	3	4	5
$L(z)$	0	0	1	3	6	10	15
	1	2	4	7	11	16	22
	2	5	8	12	17	23	30
	3	9	13	18	24	31	39
	4	14	19	25	32	40	49
	5	20	26	33	41	50	60

In Chapter 1, we saw that there are countably many finite sequences of elements from a countable set. We now give a Diophantine function $\beta(u, i)$ that explicitly enumerates all finite sequences of natural numbers. This function is due to Gödel, and it makes clever use of the Chinese remainder theorem (Theorem 3.12) in its construction.

Theorem 5.2. *For $u, i \in \mathbb{N}$, we define $\beta(u, i)$ to be the remainder when $L(u)$ is divided by $1 + (1 + i)R(u)$. Then for any finite sequence of natural numbers $a_0, \ldots, a_N$, there exists a natural number u such that $\beta(u, i) = a_i$ for $i = 0, \ldots, N$. Also, $\beta(u, i) \leq u$.*

Proof. We will make use of the Chinese remainder theorem. The first step is to give a set of $N + 1$ relatively prime numbers that we will use as our moduli. Let $y = N! \max\{a_i\}$. Then y is divisible by $1, 2, \ldots, N$, and $y \geq a_i$ for each i. The former implies that

$$1 + y,\ 1 + 2y,\ \ldots,\ 1 + (N + 1)y$$

are all relatively prime. To see this, suppose a prime p divides $1 + iy$ and $1 + jy$ for $1 \leq i < j \leq N + 1$. Then p will divide their difference, which is $(j - i)y$. Note that $j - i \leq N$. If p does not divide y, then it divides $j - i$, and hence it is at most N. However, y is divisible by

each natural number up to N, and hence $p \mid y$. Since p also divides $1 + iy$, it must divide 1, a contradiction since p is prime.

We now use the Chinese remainder theorem, which yields a number x with

$$\begin{aligned} x &\equiv a_0 \pmod{1+y} \\ x &\equiv a_1 \pmod{1+2y} \\ &\vdots \\ x &\equiv a_N \pmod{1+(N+1)y}. \end{aligned}$$

We may always take x to be a natural number, since if it were negative, we could then add to it multiples of the product of the moduli to obtain another solution. Let $u = P(x, y)$, so that $L(u) = x$ and $R(u) = y$. Then for each $i = 0, \ldots, N$ we have

$$L(u) \equiv a_i \pmod{1 + (i+1)R(u)}.$$

Since each $a_i \leq y = R(u) < 1 + (i+1)R(u)$, it follows that a_i is the remainder when $L(u)$ is divided by $1 + (i+1)R(u)$. This is our definition of $\beta(u, i)$, and hence $\beta(u, i) = a_i$ for $i = 0, \ldots, N$. We also note that $\beta(u, i) \leq L(u) \leq u$. □

It remains to show that $\beta(u, i)$ is Diophantine.

Theorem 5.3. *The function $\beta(u, i)$ of Theorem 5.2 is Diophantine.*

Proof. We claim that $z = \beta(u, i)$ holds if and only if the following three equations have a solution:

$$\begin{aligned} 2u &= (x+y)(x+y+1) + 2x, \\ x &= z + q\,(1 + (i+1)y)\,, \\ z + v + 1 &= 1 + (i+1)y. \end{aligned}$$

The first equation is equivalent to $L(u) = x$ and $R(u) = y$. The second equation is equivalent to $x \equiv z \pmod{1 + (i+1)y}$, and the third to $z < 1 + (i+1)y$. Together, these hold if and only if z is equal to the remainder when $L(u)$ is divided by $1 + (i+1)R(u)$. Thus

$z = \beta(u, i)$ if and only if there exist natural numbers q, v, x, y with

$$((x+y)(x+y+1)+2z-2u)^2+(z+q(1+(i+1)y)-x)^2 + (z+v-(i+1)y)^2 = 0.$$

Hence $\beta(u, i)$ is Diophantine. □

5.2. The Brahmagupta–Pell Equation Revisited

Our goal in the next section, Section 5.3, will be to show that the exponential function $h(b, n) = (b+1)^n$ is Diophantine. At first it may seem surprising that a set exhibiting exponential growth may be given a Diophantine definition. However, we have previously seen a particular Diophantine equation whose solutions grow exponentially: the Brahmagupta–Pell equation,

$$x^2 - dy^2 = 1,$$

of Section 3.4. Thus we revisit the equation again here. However, we will now take $d = a^2 - 1$ for a natural number a greater than 1. Then $(a, 1)$ is easily seen to be a solution. Furthermore, since it has the smallest possible y-value greater than that of the trivial solution $(1, 0)$, it is the generator. We restate some of the results of Section 3.4 in this special case.

Theorem 5.4. *All solutions (x_k, y_k) of*

$$x^2 - dy^2 = 1$$

with $d = a^2 - 1$ are given by

$$x_k + y_k\sqrt{d} = (a + \sqrt{d})^k.$$

When discussing the solutions of multiple equations, we may write $x_k = x_k(a)$ and $y_k = y_k(a)$ to distinguish between solutions from different equations. These solutions satisfy the following recursive equations:

$$x_{k\pm\ell} = x_k x_\ell \pm d y_k y_\ell, \tag{5.1}$$

$$y_{k\pm\ell} = x_\ell y_k \pm x_k y_\ell, \tag{5.2}$$

$$x_{k+1} = 2a x_k - x_{k-1}, \tag{5.3}$$

$$y_{k+1} = 2a y_k - y_{k-1}. \tag{5.4}$$

The x_k and y_k are increasing and satisfy

$$a^k \le x_k \le (2a)^k \quad \textit{and} \quad k \le y_k < (2a)^k. \tag{5.5}$$

We have $\gcd(x_k, y_k) = 1$ *and*

$$y_k \mid y_\ell \textit{ if and only if } k \mid \ell. \tag{5.6}$$

Proof. The theorem follows immediately upon letting $d = a^2 - 1$ and $(x_1, y_1) = (a, 1)$ in Theorems 3.18, 3.19, 3.20, and 3.23, and Lemmas 3.21 and 3.22. □

In the remainder of this section, we derive some new results on the solutions x_k and y_k. It may not be clear why some of these results are interesting or important, but they will be useful when we show that the y_k are Diophantine in the next section. To begin, we show some divisibility results.

Lemma 5.5. *We have $y_k^2 \mid y_{ky_k}$. Furthermore, if $y_k^2 \mid y_m$, then $y_k \mid m$.*

Proof. Since

$$\begin{aligned} x_{k\ell} + y_{k\ell}\sqrt{d} &= (a + \sqrt{d})^{k\ell} \\ &= (x_k + y_k\sqrt{d})^\ell \\ &= \sum_{i=0}^{\ell} \binom{\ell}{i} x_k^{\ell-i} y_k^i d^{i/2}, \end{aligned}$$

we have

$$y_{k\ell} = \sum_{\substack{i=0 \\ i \text{ odd}}}^{\ell} \binom{\ell}{i} x_k^{\ell-i} y_k^i d^{(i-1)/2}.$$

Reducing modulo y_k^3 leaves only the term when $i = 1$, and hence

$$y_{k\ell} \equiv \ell x_k^{\ell-1} y_k \pmod{y_k^3}. \tag{5.7}$$

Letting $\ell = y_k$ yields

$$y_k^2 \mid y_{ky_k}.$$

To show the second part of the lemma, suppose $y_k^2 \mid y_m$. Then (5.6) of Theorem 5.4 implies $k \mid m$. Writing $m = k\ell$, (5.7) yields

$$y_m \equiv \ell x_k^{\ell-1} y_k \pmod{y_k^3},$$

and hence $y_k^2 \mid \ell x_k^{\ell-1} y_k$. This implies $y_k \mid \ell x_k^{\ell-1}$. Since $\gcd(x_k, y_k) = 1$, we have $y_k \mid \ell$, and hence $y_k \mid m$. □

Our next few results are obtained from the recursive formulas (5.3) and (5.4) of Theorem 5.4. To begin, we show one more bound on y_k.

Lemma 5.6. *We have* $y_k \geq (2a-1)^{k-1}$ *for* $k \geq 1$.

Proof. Since $y_1 = 1$ and $y_2 = 2a > 2a - 1$, the result holds for $k = 1, 2$. By induction along with (5.4) of Theorem 5.4 and the fact that y_k is increasing, we have

$$\begin{aligned} y_{k+1} &= 2ay_k - y_{k-1} \\ &> 2ay_k - y_k \\ &= (2a-1)y_k \\ &\geq (2a-1)^k, \end{aligned}$$

as required. □

We now give several congruence results on the solutions to the Brahmagupta–Pell equation.

Lemma 5.7. *We have* $y_k \equiv k \pmod{a-1}$.

Proof. Since $y_k = k$ for $k = 0, 1$, the congruence holds in these cases. By induction and (5.4) of Theorem 5.4, we have

$$\begin{aligned} y_{k+1} &= 2ay_k - y_{k-1} \\ &\equiv 2k - (k-1) \pmod{a-1} \\ &\equiv k+1 \pmod{a-1}. \end{aligned}$$

We used the fact that $a \equiv 1 \pmod{a-1}$. □

Lemma 5.8. *We have*

$$x_k \equiv p^k + y_k(a-p) \pmod{2ap - p^2 - 1},$$

with the right-hand side of the congruence less than or equal to the left-hand side when $0 < p^k < a$.

Proof. For $k = 0, 1$, we have equality, and so the congruence holds. By induction with (5.3) and (5.4) of Theorem 5.4, we have

$$\begin{aligned} x_{k+1} - y_{k+1}(a-p) &= 2ax_k - x_{k-1} - (2ay_k - y_{k-1})(a-p) \\ &= 2a(x_k - y_k(a-p)) - (x_{k-1} - y_{k-1}(a-p)) \\ &\equiv 2ap^k - p^{k-1} \pmod{2ap - p^2 - 1} \\ &\equiv p^{k-1}(2ap-1) \pmod{2ap - p^2 - 1} \\ &\equiv p^{k+1} \pmod{2ap - p^2 - 1}. \end{aligned}$$

To show the inequality, we first show that

$$(a-1)y_k < x_k. \tag{5.8}$$

This is true for $k = 0$. For $k \geq 1$, since $x_k^2 - (a^2-1)y_k^2 = 1$, we have

$$\frac{x_k^2 - 1}{y_k^2} = a^2 - 1.$$

Thus

$$\frac{x_k}{y_k} > \frac{\sqrt{x_k^2 - 1}}{y_k} = \sqrt{a^2 - 1} \geq a - 1,$$

and so (5.8) holds. Now suppose $0 < p^k < a$. If $p = 1$, then (5.8) yields $1 + y_k(a-1) < 1 + x_k$, and so $1 + y_k(a-1) \leq x_k$, as required. We consider $p \geq 2$. Equality holds for $k = 0, 1$. For $k \geq 2$, we have $y_k \geq y_2 = 2a > a > p^k$. This, along with $p \geq 2$ and (5.8), yields

$$\begin{aligned} p^k + y_k(a-p) &< (1 + a - p)y_k \\ &\leq (a-1)y_k \\ &< x_k. \end{aligned}$$

This completes the proof. □

Lemma 5.9. *If* $a \equiv b \pmod{c}$, *then*

$$x_k(a) \equiv x_k(b) \pmod{c}$$

and

$$y_k(a) \equiv y_k(b) \pmod{c}$$

for all k.

Proof. Note that $x_0(a) = x_0(b) = 1$, $y_0(a) = y_0(b) = 0$, and $y_1(a) = y_1(b) = 1$. Since $x_1(a) = a$ and $x_1(b) = b$, the result holds for $k = 0$ and 1. By induction with (5.3) and (5.4) of Theorem 5.4, we have

$$\begin{aligned} x_{k+1}(a) &= 2ax_k(a) - x_{k-1}(a) \\ &\equiv 2bx_k(b) - x_{k-1}(b) \pmod{c} \\ &\equiv x_{k+1}(b) \pmod{c} \end{aligned}$$

and

$$\begin{aligned} y_{k+1}(a) &= 2ay_k(a) - y_{k-1}(a) \\ &\equiv 2by_k(b) - y_{k-1}(b) \pmod{c} \\ &\equiv y_{k+1}(b) \pmod{c}, \end{aligned}$$

completing the proof. □

Next we deduce some periodicity properties of the sequence x_k.

Lemma 5.10. *We have*

$$x_{2n\pm j} \equiv -x_j \pmod{x_n}$$

and

$$x_{4n\pm j} \equiv x_j \pmod{x_n}.$$

Proof. By (5.1) and (5.2) of Theorem 5.4, we have

$$\begin{aligned} x_{2n\pm j} &= x_n x_{n\pm j} + dy_n y_{n\pm j} \\ &= x_n x_{n\pm j} + dy_n(x_j y_n \pm x_n y_j) \\ &= x_n x_{n\pm j} \pm dx_n y_n y_j + x_j(x_n^2 - 1) \\ &\equiv -x_j \pmod{x_n}. \end{aligned}$$

Using this, we have

$$\begin{aligned} x_{4n\pm j} &= x_{2n+(2n\pm j)} \\ &\equiv -x_{2n\pm j} \pmod{x_n} \\ &\equiv x_j \pmod{x_n}. \end{aligned}$$

□

In our final result of this section, we examine the situation when $x_i \equiv x_j \pmod{x_n}$ and see what information this can give us on i and j.

Lemma 5.11. *If* $x_i \equiv x_j \pmod{x_n}$ *for* $0 < i \leq n$, *then* $j \equiv \pm i \pmod{4n}$.

Proof. First, suppose $x_i \equiv x_j \pmod{x_n}$ for $i \leq j \leq 2n$ and $n > 0$. We show that $i = j$ except in the cases $a = 2$, $n = 1$, $i = 0$, and $j = 2$. We begin with the case when x_n is odd. Set $q = (x_n - 1)/2$. Then $-q, \ldots, q$ are $2q + 1 = x_n$ consecutive integers, and hence are all distinct modulo x_n. By (5.3) of Theorem 5.4, for $n \geq 2$ we have

$$\begin{aligned} x_n - ax_{n-1} &= ax_{n-1} - x_{n-2} \\ &> x_{n-1} - x_{n-2} \\ &> 0, \end{aligned}$$

the last step holding since x_k is increasing. Hence

$$x_{n-1} < \frac{x_n}{a} \leq \frac{x_n}{2},$$

with the resulting inequality $x_{n-1} \leq x_n/2$ also holding for $n = 1$. Thus $x_{n-1} \leq q$. Since the x_k are increasing, we have

$$1 = x_0 < x_1 < \cdots < x_{n-1} \leq q.$$

By Lemma 5.10, $x_{2n-k} \equiv -x_k$ for $0 \leq k \leq n-1$. Thus modulo x_n, x_{n+1} through x_{2n} are equivalent to $-x_{n-1}$ through $-x_0 = -1$. Since

$$-q \leq -x_{n-1} < \cdots < -x_0 = -1,$$

this implies that $x_0, \ldots, x_{2n}$ are all distinct modulo x_n, which gives the result. It remains to show the result for x_n even. In this case, we set $q = x_n/2$. Then $-q + 1, \ldots, q$ are $2q = x_n$ consecutive integers and hence are all distinct modulo x_n. As before, we have

$$1 = x_0 < x_1 < \cdots < x_{n-1} \leq q$$

and

$$-q \leq -x_{n-1} < \cdots < -x_0 = -1.$$

Thus the result will hold as before unless $-q = -x_{n-1}$, whence

$$x_{n+1} \equiv -x_{n-1} \equiv -q \equiv q \equiv x_{n-1} \pmod{x_n}$$

yields a counterexample with $i = n - 1$ and $j = n + 1$. In this case, $x_n = 2q = 2x_{n-1}$. By (5.1) of Theorem 5.4, we have

$$\begin{aligned} x_n &= x_{1+(n-1)} \\ &= x_1 x_{n-1} + d y_1 y_{n-1} \\ &= a x_{n-1} + d y_{n-1}. \end{aligned}$$

Thus $(2 - a)x_{n-1} = dy_{n-1}$. Since the right-hand side is nonnegative and $a \geq 2$, we must have $a = 2$. Then $dy_{n-1} = 0$, which implies $n = 1$. Hence the result holds except when $a = 2$, $n = 1$, $i = 0$, and $j = 2$.

We now show how the lemma follows from the result that we have just proved. We assume $x_i \equiv x_j \pmod{x_n}$ with $0 < i \leq n$, and write $j = 4nq + r$ with $0 \leq r < 4n$. By Lemma 5.10,

$$x_j = x_{4nq+r} \equiv x_r \pmod{x_n}.$$

If $r \leq 2n$, then the result of the previous paragraph implies $i = r$, as the exceptional case cannot occur since it contradicts $0 < i \leq n$. If $2n < r < 4n$, then $0 < 4n - r < 2n$. By Lemma 5.10, we have

$$x_{4n-r} \equiv x_r \pmod{x_n}.$$

Again using the result of the previous paragraph, we have $i = 4n - r$, as the exceptional case cannot occur since it contradicts $i > 0$ and $4n - r > 0$. Thus we have either $i = r$ or $i = 4n - r \equiv -r \pmod{4n}$, and hence

$$j = 4nq + r \equiv r \equiv \pm i \pmod{4n},$$

which completes the proof. □

5.3. The Exponential Function Is Diophantine

In this section we will show that the function $h(b, n) = (b + 1)^n$ is Diophantine. Before we can do this, we first show that $y = y_{k+1}(a)$ can be given a Diophantine definition. Given y and $a \geq 2$, it is easy to give a Diophantine definition for y being the second component of *some* solution to the Brahmagupta–Pell equation, since this holds if and only if there exists x with $x^2 - (a^2 - 1)y^2 = 1$. However, given y, $a \geq 2$, and k, it is much more difficult to give a Diophantine

definition of y being the second component of the $(k+1)$-th solution to the Brahmagupta–Pell equation. This is what we will do now.

Theorem 5.12. *Given a, k, and y with $a \geq 2$, the system*

P1. $x^2 - (a^2-1)y^2 = 1$,

P2. $u^2 - (a^2-1)v^2 = 1$,

P3. $s^2 - (b^2-1)t^2 = 1$,

P4. $v = 4ry^2$,

P5. $b = a + u^2(u^2 - a)$,

P6. $s = x + cu$,

P7. $t = k + 1 + 4dy$,

P8. $y = k + e + 1$,

has a solution in the remaining variables if and only if $y = y_{k+1}(a)$.

Proof. We begin with the forward direction. Suppose equations P1 to P8 have a solution. If $b = 0$, then P3 implies $s = 0$ or $t = 0$. If $s = 0$, then P6 implies $x = 0$, which is impossible by P1. The case $t = 0$ is impossible by P7. If $b = 1$, then P3 implies $s = 1$, and so P6 implies $x = 0$ or $x = 1$. By P1 we cannot have $x = 0$, and so $x = 1$, which implies $y = 0$. However, this contradicts P8. Thus $b \geq 2$. By P1, P2, and P3, we have

$$\begin{gathered} x = x_i(a),\ y = y_i(a), \\ u = x_n(a),\ v = y_n(a), \\ s = x_j(b),\ \text{and } t = y_j(b) \end{gathered}$$

for some i, j, and n. Since $x \geq 2$, we must have $i \geq 1$. By P4, $y \leq v$, and so $i \leq n$. If $u = 1$, then P5 implies $b = 1$, a contradiction. Thus $x_n(a) = u \geq 2$. Equation P6 implies

$$x_j(b) \equiv x_i(a) \pmod{x_n(a)}.$$

By P5, $b \equiv a \pmod{x_n(a)}$. Thus Lemma 5.9 implies

$$x_j(a) \equiv x_j(b) \pmod{x_n(a)},$$

and hence

$$x_i(a) \equiv x_j(a) \pmod{x_n(a)}.$$

We have now met the conditions to apply Lemma 5.11, which yields $j \equiv \pm i \pmod{4n}$. By P4, $y_i(a)^2 \mid y_n(a)$. By Lemma 5.5, this yields $y_i(a) \mid n$. Hence

$$j \equiv \pm i \pmod{4y_i(a)}. \tag{5.9}$$

By P4, $v \equiv 0 \pmod{4y}$, and so P2 implies $u^2 \equiv 1 \pmod{4y}$. By P5, we then have $b \equiv 1 \pmod{4y_i(a)}$, and so $4y_i(a) \mid (b-1)$. Since Lemma 5.7 yields $(b-1) \mid (y_j(b) - j)$, we have $4y_i(a) \mid (y_j(b) - j)$, and hence

$$y_j(b) \equiv j \pmod{4y_i(a)}. \tag{5.10}$$

By P7, we have $y_j(b) \equiv k+1 \pmod{4y_i(a)}$. This with (5.9) and (5.10) yields

$$k+1 \equiv \pm i \pmod{4y_i(a)}.$$

Now P8 implies $k+1 \leq y_i(a)$, and (5.5) of Theorem 5.4 implies $i \leq y_i(a)$. Thus $k+1 = i$, and so $y = y_{k+1}(a)$, as required.

We now show the reverse direction. Suppose $y = y_{k+1}(a)$. Set $x = x_{k+1}(a)$ to satisfy P1. Let

$$m = 4(k+1)y_{k+1}(a),$$

and set $u = x_m(a)$, $v = y_m(a)$ to satisfy P2. By (5.2) of Theorem 5.4,

$$y_{2(k+1)y} = y_{(k+1)y+(k+1)y} = 2x_{(k+1)y}y_{(k+1)y},$$

and so

$$\begin{aligned} y_m &= y_{4(k+1)y} \\ &= y_{2(k+1)y+2(k+1)y} \\ &= 2x_{2(k+1)y}y_{2(k+1)y} \\ &= 4x_{(k+1)y}x_{2(k+1)y}y_{(k+1)y}. \end{aligned}$$

By Lemma 5.5, $y_{k+1}^2 \mid y_{(k+1)y}$, and so $4y_{k+1}^2 \mid y_m$. That is, $4y^2 \mid v$, and so we may choose r to satisfy P4. Set

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

We need to be sure that $b \geq 0$ in order for P5 to be satisfied, and we will in fact eventually require $b \geq 2$. Since $y = y_{k+1}(a) \geq 1$, we have

$m = 4(k+1)y_{k+1}(a) \geq 4$ and hence $u = x_m(a) > 1$ and $v = y_m(a) > 1$. By P2,

$$u^2 - 1 = (a^2 - 1)v^2 > a^2 - 1,$$

and so $a \leq a^2 < u^2 < u^2 + 1$. Multiplying by $u^2 - 1$ yields

$$a(u^2 - 1) < u^4 - 1,$$

and so

$$1 < a + u^4 - au^2 = a + u^2(u^2 - a) = b.$$

Therefore $b \geq 2$, and so P5 is satisfied. We now set $s = x_{k+1}(b)$ and $t = y_{k+1}(b)$ to satisfy P3. By (5.5), $y_{k+1}(a) \geq k+1$, and so we can choose e to satisfy P8. As seen above, $u^2 > a$, and hence $u^4 > au^2$, which implies $u^2(u^2 - a) > 0$. Thus

$$b = a + u^2(u^2 - a) > a.$$

This implies $s = x_{k+1}(b) > x_{k+1}(a) = x$. By P5, $a \equiv b \pmod{x_m(a)}$. Thus Lemma 5.9 implies

$$x_{k+1}(a) \equiv x_{k+1}(b) \pmod{x_m(a)}.$$

That is, $x \equiv s \pmod{u}$. This and the fact that $x < s$ allows us to choose c to satisfy P6. By (5.5) of Theorem 5.4,

$$k + 1 \leq y_{k+1}(b) = t.$$

By Lemma 5.7,

$$t = y_{k+1}(b) \equiv k + 1 \pmod{b - 1},$$

and so $(b-1) \mid (k+1-t)$. By P4, $v \equiv 0 \pmod{4y}$. Then by P2, we have $u^2 \equiv 1 \pmod{4y}$. This with P5 implies $b \equiv 1 \pmod{4y}$, and hence $4y \mid (b-1)$. We then have $4y \mid (k+1-t)$ and so $t \equiv k+1 \pmod{4y}$. This and the fact that $k + 1 \leq t$ allows us to choose d to satisfy P7. □

We can eliminate some of the equations and variables in P1 to P8. Using equations P4 to P7, we eliminate v, b, s, and t from the remaining equations, which yields the following theorem.

Theorem 5.13. *Given a, k, and y with $a \geq 2$, the system*

B1. $x^2 - (a^2 - 1)y^2 = 1$,

B2. $u^2 - 16(a^2 - 1)r^2y^4 = 1$,

B3. $(x + cu)^2 - ((a + u^2(u^2 - a))^2 - 1)(k + 1 + 4dy)^2 = 1,$

B4. $y = k + e + 1$

has a solution in the remaining variables if and only if $y = y_{k+1}(a)$.

We are almost ready to prove that the exponential function is Diophantine. We require one last lemma before we do so.

Lemma 5.14. *For* $e \geq 2$, *if*

$$e^3(e + 2)(a + 1)^2 + 1$$

is a perfect square, then $e - 1 + e^{e-2} \leq a$. *Also, given positive integers* e *and* t, *we can satisfy* $e^3(e + 2)(a + 1)^2 + 1 = \ell^2$ *with a value of* a *satisfying* $t \mid (a + 1)$.

Proof. To show the first part, write $m = e + 1$. We then have

$$(m - 1)^3(m + 1)(a + 1)^2 + 1 = n^2$$

for some n. Rearranging yields

$$n^2 - (m^2 - 1)\big((m - 1)(a + 1)\big)^2 = 1,$$

a Brahmagupta–Pell equation. Since $m \geq 3$, we have

$$(m - 1)(a + 1) = y_k(m)$$

for some k. Since $(m - 1)(a + 1) \geq 2$, we have $k \neq 0$. By Lemma 5.7, $(m - 1) \mid (y_k(a) - k)$. Hence $(m - 1) \mid k$. Since $k \neq 0$, we have $m - 1 \leq k$. Now for $m \geq 4$, we have $m - 2 < m - 1 \leq (m - 1)^{m-3}$. Thus $(m-2)(m-1) \leq (m-1)^{m-2}$, this inequality holding for $m = 3$ as well. Hence,

$$\begin{aligned}
(m - 2)(m - 1) + (m - 1)^{m-2} &\leq 2(m - 1)^{m-2} \\
&\leq 2^{m-2}(m - 1)^{m-2} \\
&= (2m - 2)^{m-2} \\
&< (2m - 1)^{m-2} \\
&\leq y_{m-1}(m) \qquad \text{(by Lemma 5.6)} \\
&\leq y_k(m) \qquad \text{(since } m - 1 \leq k\text{)} \\
&= (m - 1)(a + 1).
\end{aligned}$$

Dividing by $m-1$ and replacing m with $e+1$ yields $e-1+e^{e-2} < a+1$, and hence $e-1+e^{e-2} \le a$.

We now show the second part of the lemma. Since

$$(m^2-1)(m-1)^2t^2$$

is not a perfect square when $m > 1$, the Brahmagupta–Pell equation

$$x^2 - \left[(m^2-1)(m-1)^2t^2\right] y^2 = 1$$

has a nontrivial solution with $y \ge 1$. Then

$$(m-1)^3(m+1)(ty)^2 + 1 = x^2.$$

Since $t \ge 1$ and $y \ge 1$, we have $ty \ge 1$. Taking $m = e+1$, $a+1 = ty$, and $\ell = x$ yields

$$e^3(e+2)(a+1)^2 + 1 = \ell^2$$

with $t \mid (a+1)$. □

We may now give a Diophantine definition of the exponential function.

Theorem 5.15. *Given numbers f, x, and n with $x \ge 1$, the system*

E1. $v^2 - (a^2-1)w^2 = 1$,

E2. $u^2 - 16(a^2-1)r^2w^4 = 1$,

E3. $(v+cu)^2 - ((a+u^2(u^2-a))^2 - 1)(n+1+4dw)^2 = 1$,

E4. $w = n+p+1$,

E5. $e^3(e+2)(a+1)^2 + 1 = \ell^2$,

E6. $e = x+n+f+2$,

E7. $v = f + w(a-x) + i(2ax - x^2 - 1)$

has a solution in the remaining variables if and only if $f = x^{n+1}$.

Proof. We show the forward direction first. Suppose equations E1 to E7 are satisfied. Equation E6 implies $e \ge 2$, and so Lemma 5.14 with E5 implies $e - 1 + e^{e-2} \le a$. With E6, this implies

$$x+n+f+1+(x+n+f+2)^{x+n+f} \le a.$$

Hence, we have $f < a$ and $x^{n+1} < a$. It also implies that $a \geq 2$. Since $1 \leq x < a$, we have

$$x^2 < ax = 2ax - ax \leq 2ax - a.$$

Thus $x^2 \leq 2ax - a - 1$, and so $a \leq 2ax - x^2 - 1$. From this it follows that $f < 2ax - x^2 - 1$ and $x^{n+1} < 2ax - x^2 - 1$. By equations B1 to B4 of Theorem 5.13, equations E1 to E4 imply $v = x_{n+1}(a)$ and $w = y_{n+1}(a)$. By Lemma 5.8, it follows that

$$v \equiv x^{n+1} + w(a - x) \pmod{2ax - x^2 - 1}.$$

Equation E7 yields

$$v \equiv f + w(a - x) \pmod{2ax - x^2 - 1}.$$

Hence $f \equiv x^{n+1} \pmod{2ax - x^2 - 1}$. The inequalities given above then imply $f = x^{n+1}$, as required.

We now show the reverse direction. Suppose $f = x^{n+1}$ for $x \geq 1$. Set $e = x + n + f + 2$ to satisfy E6. Lemma 5.14 implies that we can find a and ℓ to satisfy E5 with $a \geq 2$, since we can have $3 \mid (a + 1)$. Set $w = y_{n+1}(a)$. Then Theorem 5.13 implies that we may find c, d, p, r, u, and v satisfying E1 to E4, and that $v = x_{n+1}(a)$. Lemma 5.8 then implies

$$v \equiv f + w(a - x) \pmod{2ax - x^2 - 1}$$

with the right-hand side less than or equal to the left-hand side. Thus we may choose i to satisfy E7. □

Theorem 5.16. *The function $h(b, n) = (b + 1)^n$ is Diophantine.*

Proof. We would like to use Theorem 5.15, but we would like to allow the exponent to be equal to 0. A Diophantine definition can be given for $g = x^n$ with $x \geq 1$ by appending an additional equation to those of Theorem 5.15:

E8. $xg = f$.

Thus $h(b, n) = (b+1)^n$ can be shown to be Diophantine by replacing each x with $b + 1$ in E1 to E8, moving all terms to one side in each equation, and then summing their squares. □

5.4. More Diophantine Functions

We previously asked if the powers of 2 form a Diophantine set. Given the results of the previous section, we now see that the answer to that question is yes. What about the prime numbers? To answer this question, we require Diophantine definitions for a few additional functions, including the binomial coefficients and the factorial function.

We give a Diophantine definition of the binomial coefficients. The basic idea is to use the binomial theorem on $(u+1)^a$ and to take u large enough so that the coefficients of the powers of u in the expansion are the digits of the base u representation of $(u+1)^a$.

Theorem 5.17. *Given numbers a, b, and c with $b \leq a$, the system*

C1. $u > 2^a$,

C2. $(u+1)^a = xu^{b+1} + cu^b + y$,

C3. $c < u$,

C4. $y < u^b$

has a solution in the remaining variables if and only if $c = \binom{a}{b}$. Thus the function $f(a,b) = \binom{a}{b}$ is Diophantine.

Proof. We show the forward direction. Suppose equations C1 to C4 are satisfied. Then for $0 \leq b \leq a$, C1 implies

$$\binom{a}{b} \leq \sum_{i=0}^{a} \binom{a}{i} = (1+1)^a = 2^a < u.$$

Since

$$(u+1)^a = \sum_{i=0}^{a} \binom{a}{i} u^i$$

and $b \leq a$, the above inequality implies that $\binom{a}{b}$ is the bth digit in the base u representation of $(u+1)^a$. However, C2 with the inequalities in C3 and C4 imply that c is the bth digit in the base u representation of $(u+1)^a$. Thus $c = \binom{a}{b}$, as required.

We now show the converse. Suppose that $c = \binom{a}{b}$ with $b \leq a$. We can choose any $u > 2^a$ to satisfy C1. Then, as above, $c = \binom{a}{b}$ is the bth digit in the base u representation of $(u+1)^a$. Thus $c < u$,

satisfying C3, and we may find numbers x and y to satisfy C2 and C4.

Since the exponential function and the less than relation are Diophantine, it follows that the function $f(a,b) = \binom{a}{b}$ is Diophantine. □

We now give a Diophantine definition of the factorial function.

Theorem 5.18. *Given numbers n and b with $n, b \geq 1$, the system*

F1. $v = 2b$,

F2. $a = v^b$,

F3. $c = \binom{a}{b}$,

F4. $f = a^b$,

F5. $nc \leq f$,

F6. $f < (n+1)c$

has a solution in the remaining variables if and only if $n = b!$. Thus the function $g(b) = b!$ is Diophantine.

Proof. We claim that for $a = (2b)^b$,

$$b! \leq \frac{a^b}{\binom{a}{b}} < b! + 1.$$

The result holds for $b = 1$, with equality in the first inequality. We assume $b \geq 2$. Then

$$\begin{aligned} \frac{a^b}{\binom{a}{b}} &= \frac{a^b b!}{a(a-1)(a-2)\cdots(a-(b-1))} \\ &= \frac{b!}{\left(1-\frac{1}{a}\right)\left(1-\frac{2}{a}\right)\cdots\left(1-\frac{b-1}{a}\right)}. \end{aligned}$$

On one hand,

$$\left(1-\frac{1}{a}\right)\left(1-\frac{2}{a}\right)\cdots\left(1-\frac{b-1}{a}\right) < 1,$$

and so

$$\frac{a^b}{\binom{a}{b}} > b!.$$

On the other hand,

$$\begin{aligned}
\frac{1}{\left(1-\frac{1}{a}\right)\left(1-\frac{2}{a}\right)\cdots\left(1-\frac{b-1}{a}\right)} &< \frac{1}{\left(1-\frac{b}{a}\right)^{b-1}} \\
&= \left[1+\frac{b}{a}+\left(\frac{b}{a}\right)^2+\cdots\right]^{b-1} \\
&\le \left[1+\frac{b}{a}\left(1+\frac{1}{2}+\left(\frac{1}{2}\right)^2+\cdots\right)\right]^{b-1} \\
&= \left(1+\frac{2b}{a}\right)^{b-1} \\
&= \sum_{j=0}^{b-1}\binom{b-1}{j}\left(\frac{2b}{a}\right)^j \\
&= 1+\frac{2b}{a}\sum_{j=1}^{b-1}\binom{b-1}{j}\left(\frac{2b}{a}\right)^{j-1} \\
&< 1+\frac{2b}{a}\sum_{j=0}^{b-1}\binom{b-1}{j} \\
&= 1+\frac{2b}{a}2^{b-1} \\
&= 1+\frac{1}{b^{b-1}} \\
&\le 1+\frac{1}{b!}.
\end{aligned}$$

In both finite sums, we used the fact that $\frac{b}{a} < \frac{2b}{a} \le 1$, which holds for our choice of a. Thus we have

$$\frac{a^b}{\binom{a}{b}} < b!\left(1+\frac{1}{b!}\right) = b!+1,$$

which completes the proof of the claim.

Now suppose equations F1 to F6 hold. This implies

$$n \le \frac{a^b}{\binom{a}{b}} < n+1.$$

With the above claim, this yields

$$b! - 1 \leq \frac{a^b}{\binom{a}{b}} - 1 < n \leq \frac{a^b}{\binom{a}{b}} < b! + 1,$$

and so $b! - 1 < n < b! + 1$. Since these are natural numbers, we must have $n = b!$. Conversely, if $n = b!$, then the above claim yields

$$\binom{a}{b} n \leq a^b < (n+1)\binom{a}{b}$$

for $a = (2b)^b$. Thus we may choose the remaining variables to satisfy equations F1 through F6.

Since the exponential function, the binomial coefficient function, and the less than relation are Diophantine, it follows that the function $g(b) = b!$ is Diophantine. □

We note that we may now use Wilson's theorem (Theorem 3.11) to give a Diophantine definition of the prime numbers. Wilson's theorem states that $n > 1$ is prime if and only if $(n-1)! \equiv -1 \pmod{n}$. That is, $n > 1$ is prime if and only if there exist m, f, and k with $m + 1 = n$, $f = m!$, and $f + 1 = kp$. Since we have a Diophantine definition for the factorial function, this implies that the set of prime numbers is Diophantine. Putnam's result then implies the existence of a polynomial whose nonnegative range is precisely the set of prime numbers! In fact, we will work to explicitly write down such a polynomial in the next chapter.

We will require one more Diophantine definition for a certain product function, namely

$$\prod_{k=1}^{y} (a + bk).$$

Note that when setting $a = 0$ and $b = 1$, this product is equal to $y!$. Thus this function is a generalization of the factorial function.

Theorem 5.19. *Given a, b, y, and z with $y, b \geq 1$, the system*

M1. $r = a + by$,

M2. $s = r^y$,

M3. $M = bs + 1$,

M4. $bq = a + Mt$,
M5. $u = b^y$,
M6. $v = y!$,
M7. $z < M$,
M8. $w = q + y$,
M9. $x = \binom{w}{y}$,
M10. $z + Mp = uvx$

has a solution in the remaining variables if and only if

$$z = \prod_{k=1}^{y}(a + bk).$$

Proof. We first note that if $bq \equiv a \pmod{M}$, then

$$\begin{aligned}\prod_{k=1}^{y}(a+bk) &\equiv \prod_{k=1}^{y}(bq+bk) \pmod{M}\\ &\equiv b^y \prod_{k=1}^{y}(q+k) \pmod{M}\\ &\equiv b^y y!\binom{q+y}{y} \pmod{M}.\end{aligned} \tag{5.11}$$

We first show the forward direction. Suppose equations M1 to M10 are satisfied. Then M1 to M3 imply $M = b(a+by)^y + 1$. Hence,

$$M > (a+by)^y = \prod_{k=1}^{y}(a+by) > \prod_{k=1}^{y}(a+bk).$$

By M4 we have $bq \equiv a \pmod{M}$, and so (5.11) holds. However, M10 together with M5, M6, M8, and M9 imply

$$z \equiv b^y y!\binom{q+y}{y} \pmod{M}.$$

Since

$$z \equiv \prod_{k=1}^{y}(a+bk) \pmod{M}$$

with, by M7, both sides less than M, we have $z = \prod_{k=1}^{y}(a+bk)$, as required.

We now show the converse. Suppose that

$$z = \prod_{k=1}^{y} (a + bk)$$

with $b \geq 1$. We can then set r to satisfy M1, s to satisfy M2, and M to satisfy M3. It follows that $M = b(a + by)^y + 1$, and so

$$M > \prod_{k=1}^{y} (a + bk) = z,$$

as above, satisfying M7. We also have that $\gcd(M, b) = 1$. Thus we may find q with $bq \equiv a \pmod{M}$. Taking q large enough so that $bq > a$, we may find t to satisfy M4. We may set u, v, w, and x to satisfy M5, M6, M8, and M9. By (5.11), we have

$$z \equiv b^y y! \binom{q+y}{y} \pmod{M}.$$

Since $a < bq$, it follows that

$$z = \prod_{k=1}^{y} (a + bk) < \prod_{k=1}^{y} (bq + bk) = b^y y! \binom{q+y}{y},$$

and so p may be found to satisfy M10. □

5.5. The Bounded Universal Quantifier

We have previously seen that using conjunction, disjunction, and the existential quantifier on Diophantine expressions will yield another Diophantine expression. What about the remaining logical connectives and quantifiers: negation, implication, and the universal quantifier? We will eventually see that the use of these on Diophantine expressions can yield expressions that are not Diophantine. This is unfortunate, as the universal quantifier would be very useful to have. Happily, we have something slightly weaker that may be used: the bounded universal quantifier "for all y with $y \leq x$," written $(\forall y)_{\leq x}$.

The bulk of the work in showing that the bounded universal quantifier may be used is contained in the following lemma, where a bounded universal quantifier is replaced with a collection of Diophantine expressions.

Lemma 5.20. *Given y, u, $x_1, \ldots, x_n$, and a polynomial*

$$P(y, k, x_1, \ldots, x_n, y_1, \ldots, y_m),$$

let $Q(y, u, x_1, \ldots, x_n)$ be a polynomial satisfying the following properties:

(a) $Q(y, u, x_1, \ldots, x_n) \geq u$,

(b) $Q(y, u, x_1, \ldots, x_n) \geq y$, *and*

(c) *when $k \leq y$ and $y_1, \ldots, y_m \leq u$, we have*

$$|P(y, k, x_1, \ldots, x_n, y_1, \ldots, y_m)| \leq Q(y, u, x_1, \ldots, x_n).$$

Then

$$(\forall k)_{\leq y}(\exists y_1, \ldots, y_m)_{\leq u}(P(y, k, x_1, \ldots, x_n, y_1, \ldots, y_m) = 0)$$

if and only if there exist c, t, $a_1, \ldots, a_m$ with

(1) $1 + (c+1)t = \prod_{k=0}^{y}(1 + (k+1)t)$,

(2) $t = Q(y, u, x_1, \ldots, x_n)!$,

(3) *for each i with $1 \leq i \leq m$, we have*

$$(1 + (c+1)t) \mid \prod_{j=0}^{u}(a_i - j),$$

and

(4) $P(y, c, x_1, \ldots, x_n, a_1, \ldots, a_m) \equiv 0 \pmod{1 + (c+1)t}$.

Proof. We first prove the reverse direction. Suppose equations (1) to (4) are satisfied. For each $k = 0, \ldots, y$, let p_k be a prime dividing $1 + (k+1)t$. Note that since (2) implies $t \geq 1$, $1 + (k+1)t$ will always have a prime divisor. For each p_k with $0 \leq k \leq y$ and each i with $1 \leq i \leq m$, let $y_{i,k}$ be the remainder when a_i is divided by p_k. We will show that these $y_{i,k}$ are all bounded above by u and satisfy P, which will complete this direction of the proof.

For each k we have

$$p_k \mid (1 + (k+1)t),$$

$$(1 + (k+1)t) \mid (1 + (c+1)t),$$

and, for each i,

$$(1 + (c + 1)t) \mid \prod_{j=0}^{u} (a_i - j).$$

Thus for any k and i, we have $p_k \mid \prod_{j=0}^{u}(a_i - j)$. Since p_k is prime, for some particular j with $0 \leq j \leq u$, we have $p_k \mid (a_i - j)$. Then

$$j \equiv a_i \equiv y_{i,k} \pmod{p_k}.$$

Since p_k is a nontrivial divisor of $1+(k+1)t$, it follows that p_k cannot divide t. All numbers up to $Q(y, u, x_1, \ldots, x_n)$ divide

$$t = Q(y, u, x_1, \ldots, x_n)!,$$

and so we have $p_k > Q(y, u, x_1, \ldots, x_n)$. Thus (a) yields $p_k > u$, and hence $0 \leq j \leq u < p_k$. Since $y_{i,k}$ is a remainder upon division by p_k, we have $0 \leq y_{i,k} < p_k$. Thus $y_{i,k} \equiv j \pmod{p_k}$ implies $j = y_{i,k}$, and hence $y_{i,k} \leq u$, as required.

Since p_k divides both $1 + (c + 1)t$ and $1 + (k + 1)t$, they are congruent to 0, and hence to each other modulo p_k. We have seen that p_k does not divide t, and so t is invertible modulo p_k. Thus $k \equiv c \pmod{p_k}$. We know $y_{i,k} \equiv a_i \pmod{p_k}$, and so

$$\begin{aligned} &P(y, k, x_1, \ldots, x_n, y_{1,k}, \ldots, y_{m,k}) \\ &\qquad \equiv P(y, c, x_1, \ldots, x_n, a_1, \ldots, a_m) \pmod{p_k}. \end{aligned}$$

Since $p_k \mid (1+(c+1)t)$, (4) implies $P(y, c, x_1, \ldots, x_n, a_1, \ldots, a_m) \equiv 0 \pmod{p_k}$, and hence

$$P(y, k, x_1, \ldots, x_n, y_{1,k}, \ldots, y_{m,k}) \equiv 0 \pmod{p_k}.$$

By (c) we have that

$$|P(y, k, x_1, \ldots, x_n, y_{1,k}, \ldots, y_{m,k})| \leq Q(y, u, x_1, \ldots, x_n) < p_k.$$

Therefore, $P(y, k, x_1, \ldots, x_n, y_{1,k}, \ldots, y_{m,k}) = 0$. This completes the proof of the reverse direction.

For the forward direction, suppose for each $k = 0, \ldots, y$ we can find $y_{i,k}$ for $1 \leq i \leq m$ with $0 \leq y_{i,k} \leq u$ and

$$P(y, k, x_1, \ldots, x_n, y_{1,k}, \ldots, y_{m,k}) = 0.$$

Set $t = Q(y, u, x_1, \ldots, x_n)!$ to satisfy (2). Since

$$\prod_{k=0}^{y}(1 + (k+1)t) \equiv 1 \pmod{t},$$

there is a c with

$$1 + (c+1)t = \prod_{k=0}^{y}(1 + (k+1)t).$$

The fact that $t \geq 1$ implies $c \geq 0$, and thus (1) is satisfied.

We show that the numbers $1 + (k+1)t$ for $k = 0, \ldots, y$ are relatively prime. To see this, suppose p is a prime with $p \mid (1+(k+1)t)$ and $p \mid (1+(\ell+1)t)$ for some $0 \leq k < \ell \leq y$. Then p will divide their difference $(\ell - k)t$. Note that p does not divide t. Thus $p \mid (\ell - k)$, and so $p \leq y$. Then (b) implies $p \leq Q(y, u, x_1, \ldots, x_n)$. Since $t = Q(y, u, x_1, \ldots, x_n)!$, it follows that $p \mid t$, a contradiction. Thus the numbers $1 + (k+1)t$ for $k = 0, \ldots, y$ are relatively prime, and hence form an admissible sequence of moduli for the Chinese remainder theorem. Hence, for each i we may find a number a_i such that

$$a_i \equiv y_{i,k} \pmod{1 + (k+1)t} \quad \text{for } k = 0, \ldots, y.$$

We note that we may take $a_i > u$ since adding a multiple of

$$\prod_{k=0}^{y}(1 + (k+1)t) = 1 + (c+1)t$$

to a_i will yield the same congruences. Since $(1+(k+1)t) \mid (1+(c+1)t)$, it follows that $(1 + (k+1)t)$ divides

$$1 + (k+1)t - (1 + (c+1)t) = (k - c)t.$$

Hence,

$$(k - c)t \equiv 0 \pmod{1 + (k+1)t}.$$

Since $\gcd(t, 1+(k+1)t) = 1$, t is invertible modulo $1+(k+1)t$. This implies

$$k \equiv c \pmod{1 + (k+1)t}.$$

Thus

$$\begin{aligned} &P(y, c, x_1, \ldots, x_n, a_1, \ldots, a_m) \\ &\qquad \equiv P(y, k, x_1, \ldots, x_n, y_{1,k}, \ldots, y_{m,k}) \pmod{1 + (k+1)t} \\ &\qquad \equiv 0 \pmod{1 + (k+1)t}. \end{aligned}$$

Since $1+(k+1)t$ for $k = 0, \ldots, y$ are relatively prime and each divides $P(y, c, x_1, \ldots, x_n, a_1, \ldots, a_m)$, so does their product. Hence,

$$P(y, c, x_1, \ldots, x_n, a_1, \ldots, a_m) \equiv 0 \pmod{1 + (c+1)t},$$

satisfying (4). Since $a_i \equiv y_{i,k} \pmod{1 + (k+1)t}$, we have

$$(1 + (k+1)t) \mid (a_i - y_{i,k}).$$

As $0 \leq y_{i,k} \leq u$, it follows that

$$(1 + (k+1)t) \Big| \prod_{j=0}^{u} (a_i - j).$$

Finally, since $1 + (k+1)t$ for $k = 0, \ldots, y$ are relatively prime, we have

$$(1 + (c+1)t) \Big| \prod_{j=0}^{u} (a_i - j).$$

and so (3) may be satisfied. □

We now show that the use of the bounded universal quantifier on a Diophantine expression yields a Diophantine expression.

Theorem 5.21. *Given y, $x_1, \ldots, x_n$, and a polynomial*

$$P(y, k, x_1, \ldots, x_n, y_1, \ldots, y_m),$$

$$(\forall k)_{\leq y} (\exists y_1, \ldots, y_m) \left(P(y, k, x_1, \ldots, x_n, y_1, \ldots, y_m) = 0\right)$$

if and only if there exist $u, c, t, a_1, \ldots, a_m, e, f, g_1, \ldots, g_m, h_1, \ldots, h_m$, and $Q(y, u, x_1, \ldots, x_n)$ with Q satisfying properties (a), (b), and (c) of Lemma 5.20 *and*

U1. $e = 1 + (c+1)t$,

U2. $e = \displaystyle\prod_{k=0}^{y} (1 + (k+1)t)$,

U3. $f = Q(y, u, x_1, \ldots, x_n)$,

U4. $t = f!$,

U5. $g_1 = a_1 - u,\ g_2 = a_2 - u,\ \ldots,\ g_m = a_m - u,$

U6. $h_1 = \prod_{k=0}^{u}(g_1 + k),\ \ldots,\ h_m = \prod_{k=0}^{u}(g_m + k),$

U7. $e \mid h_1,\ \ldots,\ e \mid h_m,$

U8. $\ell = P(y, c, x_1, \ldots, x_n, a_1, \ldots, a_m),$ *and*

U9. $e \mid \ell.$

Proof. First note that we have

$$(\forall k)_{\leq y}(\exists y_1, \ldots, y_m)[P(y, k, x_1, \ldots, x_n, y_1, \ldots, y_m) = 0]$$
$$\Longleftrightarrow$$
$$(\exists u)(\forall k)_{\leq y}(\exists y_1, \ldots, y_m)_{\leq u}[P(y, k, x_1, \ldots, x_n, y_1, \ldots, y_m) = 0].$$

The reverse implication is trivial. In the forward direction, for each $k = 0, \ldots, y$ we have numbers $y_{i,k}$ for which

$$P(y, k, x_1, \ldots, x_n, y_{1,k}, \ldots, y_{m,k}) = 0.$$

Setting u to be the maximum of the finite set

$$\{y_{i,k} : 0 \leq k \leq y, 1 \leq i \leq m\}$$

yields the result.

If we are given $P(y, k, x_1, \ldots, x_n, y_1, \ldots, y_m)$, then we may always find $Q(y, u, x_1, \ldots, x_n)$ satisfying properties (a), (b), and (c) of Lemma 5.20. For example, replacing each coefficient of P with its absolute value, setting $k = y$ and $y_i = u$ for each i, and adding u and y to the result will yield a valid polynomial Q. Thus the theorem follows from Lemma 5.20. □

Most of equations U1 to U9 of Theorem 5.21 are clearly Diophantine expressions. For U2, we may re-index as $e = \prod_{k=1}^{y+1}(1 + kt)$, satisfying the requirements of Theorem 5.19. Recall that we have each $a_i > u$, so that $g_i \geq 1$. Thus U6 may be re-indexed as $h_i = \prod_{k=1}^{u+1}((g_i - 1) + k)$, satisfying the requirements of Theorem 5.19.

We may now use the bounded universal quantifier to give a Diophantine definition for the set of primes that does not use Wilson's theorem. A number n is prime if and only if $n > 1$ and for all $x, y \leq n$, either $xy < n$, $xy > n$, $x = 1$, or $y = 1$. Since this is made up of

Diophantine expressions, disjunctions, a conjunction, and bounded universal quantifiers, it is also Diophantine.

5.6. Recursive Functions Revisited

We now have some rather strong tools at our disposal for demonstrating that a function or set is Diophantine. In this section, we will see just how far these methods may be pushed. In fact, we will show that a function is Diophantine if and only if it is recursive, and that a set is Diophantine if and only if it is computably enumerable. These notions were introduced in Section 4.2.

We recall the definition of the recursive functions. The constant function $C_0(x) = 0$, the successor function $S(x) = x + 1$, and the projection functions $P_i^n(x_1, \ldots, x_n) = x_i$ (for $n \geq 1$ and $1 \leq i \leq n$) are defined to be recursive. Additional recursive functions may be obtained by repeatedly applying to these functions the operations of composition, primitive recursion, and minimalization. The composition of given function $f(t_1, \ldots, t_m)$ with given functions $g_1(x_1, \ldots, x_n), \ldots, g_m(x_1, \ldots, x_n)$ is the function

$$h(x_1, \ldots, x_n) = f(g_1(x_1, \ldots, x_n), \ldots, g_m(x_1, \ldots, x_n)).$$

Given functions $f(x_1, \ldots, x_n)$ and $g(t_1, \ldots, t_{n+2})$, primitive recursion yields the function $h(x_1, \ldots, x_n, z)$ satisfying

$$h(x_1, \ldots, x_n, 0) = f(x_1, \ldots, x_n), \text{ and}$$
$$h(x_1, \ldots, x_n, t+1) = g(t, h(x_1, \ldots, x_n, t), x_1, \ldots, x_n).$$

Given a function $f(z, x_1, \ldots, x_n)$, minimalization yields the partial function

$$h(x_1, \ldots, x_n) = \min_z \{f(z, x_1, \ldots, x_n) = 0\}.$$

If no such z exists, then h is undefined. In order for h to be recursive, we require h to be a total function.

In Section 4.2, we defined the addition function $A(x, y) = x + y$, and the multiplication function $M(x, y) = xy$. We defined functions $\text{Quo}(a, b)$ and $\text{Rem}(a, b)$ to be the quotient and remainder, respectively, of a when divided by $b + 1$. All of these functions were shown to be recursive. We now show that some additional important functions are recursive.

Theorem 5.22. *Cantor's pairing function $P(x, y)$ and Gödel's β function $\beta(u, i)$ are recursive.*

Proof. We have

$$\begin{aligned} P(x,y) &= \frac{(x+y)(x+y+1)}{2} + x \\ &= \mathrm{Quo}((x+y)(x+y+1),1) + x \\ &= A(\mathrm{Quo}(M(A(x,y),A(x,S(y))),1),x), \end{aligned}$$

and hence $P(x, y)$ is recursive. Now, $L(z)$ is the unique x such that $P(x, y) = z$ for some y. In fact, y must then be $R(z)$, and so $y \leq z$. Thus we can characterize $L(z)$ as the minimum x such that $|P(x, y) - z| = 0$ for some $y \leq z$. Since a product is equal to zero if and only if at least one of the factors is zero, we have

$$L(z) = \min_x \left\{ \prod_{y=0}^{z} |P(x,y) - z| = 0 \right\}.$$

Similarly,

$$R(z) = \min_y \left\{ \prod_{x=0}^{z} |P(x,y) - z| = 0 \right\}.$$

In the exercises for Chapter 4, it was shown that the finite product of recursive functions is recursive. In the same chapter, we saw that the function $|x - y|$ is recursive. Thus $L(z)$ and $R(z)$ are recursive functions. Finally, $\beta(u, i)$ was defined to be the remainder when $L(u)$ is divided by $1 + (i + 1)R(u)$. That is,

$$\begin{aligned} \beta(u,i) &= \mathrm{Rem}(L(u),(i+1)R(u)) \\ &= \mathrm{Rem}(L(u),M(S(i),R(u))), \end{aligned}$$

and hence $\beta(u, i)$ is recursive. □

We now show one direction of the equivalence between Diophantine and recursive functions.

Theorem 5.23. *If $f(x_1, \ldots, x_n)$ is a Diophantine function, then it is recursive.*

Proof. Recall that f is Diophantine if and only if

$$\{(x_1, \ldots, x_n, y) : y = f(x_1, \ldots, x_n)\}$$

is a Diophantine set. Thus we have a polynomial P with integer coefficients such that $y = f(x_1, \ldots, x_n)$ if and only if there exist $t_1, \ldots, t_m$ for which $P(x_1, \ldots, x_n, y, t_1, \ldots, t_m) = 0$. Grouping positive and negative coefficients, we have $y = f(x_1, \ldots, x_n)$ if and only if there exist $t_1, \ldots, t_m$ such that

$$Q_{\text{pos}}(x_1, \ldots, x_n, y, t_1, \ldots, t_m) - Q_{\text{neg}}(x_1, \ldots, x_n, y, t_1, \ldots, t_m) = 0,$$

where Q_{pos} and Q_{neg} are polynomials with natural number coefficients. We may now choose u so that $\beta(u, 0) = y$, $\beta(u, 1) = t_1$, $\ldots$, $\beta(u, m) = t_m$. Then $y = f(x_1, \ldots, x_n)$ if and only if there exists u with

$$\begin{aligned}\big|&Q_{\text{pos}}(x_1, \ldots, x_n, \beta(u, 0), \beta(u, 1), \ldots, \beta(u, m)) \\ &\quad - Q_{\text{neg}}(x_1, \ldots, x_n, \beta(u, 0), \beta(u, 1), \ldots, \beta(u, m))\big| = 0.\end{aligned}$$

Since $y = \beta(u, 0)$, we see that $f(x_1, \ldots, x_n)$ is equal to

$$\begin{aligned}\beta\Big(\min_u \Big\{ &\big|Q_{\text{pos}}(x_1, \ldots, x_n, \beta(u, 0), \beta(u, 1), \ldots, \beta(u, m)) \\ &\quad - Q_{\text{neg}}(x_1, \ldots, x_n, \beta(u, 0), \beta(u, 1), \ldots, \beta(u, m))\big| = 0 \Big\}, 0 \Big).\end{aligned}$$

In Section 4.2, we saw that polynomials with natural number coefficients are recursive, and so Q_{pos} and Q_{neg} are recursive. We have just seen that $\beta(u, i)$ is recursive, and in Section 4.2 we saw that $|x - y|$ is recursive. It follows that f is recursive. □

We now show the converse of the previous theorem.

Theorem 5.24. *If a function is recursive, then it is Diophantine.*

Proof. It is trivial to check that the constant function $C_0(x)$, successor function $S(x)$, and projection functions $P_i^n(x_1, \ldots, x_n)$ are Diophantine. Thus we must show that the operations of composition, primitive recursion, and minimalization performed on Diophantine functions will yield another Diophantine function.

Suppose

$$h(x_1, \ldots, x_n) = f(g_1(x_1, \ldots, x_n), \ldots, g_m(x_1, \ldots, x_n))$$

is the composition of Diophantine functions $g_1, \ldots, g_m$ with Diophantine function f. Then $y = h(x_1, \ldots, x_n)$ if and only if there

exist $t_1, \ldots, t_m$ with $t_1 = g_1(x_1, \ldots, x_n)$, $t_2 = g_2(x_1, \ldots, x_n)$, $\ldots$, $t_m = g_m(x_1, \ldots, x_n)$, and $y = f(t_1, \ldots, t_m)$. Since this is the conjunction of Diophantine expressions, h is Diophantine.

Suppose $h(x_1, \ldots, x_n, z)$ is obtained through primitive recursion on Diophantine functions $f(x_1, \ldots, x_n)$ and $g(t_1, \ldots, t_{n+2})$. That is, suppose $h(x_1, \ldots, x_n, 0) = f(x_1, \ldots, x_n)$ and

$$h(x_1, \ldots, x_n, t+1) = g(t, h(x_1, \ldots, x_n, t), x_1, \ldots, x_n).$$

We use Gödel's β function to encode

$$h(x_1, \ldots, x_n, 0), \ldots, h(x_1, \ldots, x_n, z)$$

as follows: there is some u with

$$\beta(u, 0) = h(x_1, \ldots, x_n, 0) = f(x_1, \ldots, x_n)$$

and for $t + 1 \leq z$,

$$\begin{aligned} \beta(u, t+1) &= h(x_1, \ldots, x_n, t+1) \\ &= g(t, h(x_1, \ldots, x_n, t), x_1, \ldots, x_n) \\ &= g(t, \beta(u, t), x_1, \ldots, x_n). \end{aligned}$$

Hence $y = h(x_1, \ldots, x_n, z)$ if and only if there exists u such that the following three statements hold:

(a) $(\exists w)(w = \beta(u, 0)$ and $w = f(x_1, \ldots, x_n))$,

(b) $(\forall t)_{\leq z}\Big((t = z)$ or $(\exists w)(w = \beta(u, t+1)$
and $w = g(t, \beta(u, t), x_1, \ldots, x_n))\Big)$,

(c) $y = \beta(u, z)$.

In Theorem 5.3, we showed that $\beta(u, i)$ is Diophantine. Since the conjunction of the above three statements is the conjunction, disjunction, existential quantification, and bounded universal quantification of Diophantine expressions, it too is Diophantine. Thus h is Diophantine.

Finally, suppose

$$h(x_1, \ldots, x_n) = \min_y \{f(y, x_1, \ldots, x_n) = 0\}$$

with f Diophantine and h total. Then $y = h(x_1, \ldots, x_n)$ if and only if $f(y, x_1, \ldots, x_n) = 0$ and

$$(\forall t)_{\leq y} \left[(t = y) \text{ or } f(x_1, \ldots, x_n, t) > 0 \right].$$

Since this is the conjunction, disjunction, and bounded universal quantification of Diophantine expressions, it is also Diophantine. Thus h is Diophantine. □

In summary, we have shown that a function is Diophantine if and only if it is recursive. As discussed in Chapter 4, the recursive functions are the same as the computable functions. Thus our definition of a Diophantine function is really just another way to characterize the class of computable functions!

We now give the link between Diophantine sets and computably enumerable sets.

Theorem 5.25. *A set of natural numbers is Diophantine if and only if it is computably enumerable.*

Proof. Suppose S is a computably enumerable set of natural numbers. Then S is equal to the range of a computable function f. That is, $y \in S$ if and only if there exist $x_1, \ldots, x_n$ such that $y = f(x_1, \ldots, x_n)$. Since we now know that computable functions are Diophantine, we have $y \in S$ if and only if there exist $x_1, \ldots, x_n, t_1, \ldots, t_m$ with $P(x_1, \ldots, x_n, y, t_1, \ldots, t_m) = 0$. This implies that S is Diophantine.

For the converse, we make use of the Church–Turing thesis. Suppose S is Diophantine. Then $y \in S$ if and only if there exist $t_1, \ldots, t_n$ with $P(y, t_1, \ldots, t_m) = 0$. Using the inverse of Cantor's pairing function repeatedly, we may traverse through all $(m+1)$-tuples of natural numbers. For each $(m+1)$-tuple $(y, t_1, \ldots, t_m)$, we determine if $P(y, t_1, \ldots, t_m) = 0$. If so, we list y. Thus we have an algorithm to list the members of S. By the Church–Turing thesis, this implies S is computably enumerable. □

5.7. Solution of Hilbert's Tenth Problem

We have now seen enough to quickly resolve Hilbert's tenth problem.

Theorem 5.26 (Negative solution to Hilbert's tenth problem)**.** *There does not exist an algorithm that will determine if an arbitrary Diophantine equation has natural number solutions.*

Proof. In Theorem 4.2, we described a set U of natural numbers that is computably enumerable but not decidable. Since U is computably enumerable, it is Diophantine. Thus $x \in U$ if and only if there exist $y_1, \ldots, y_m$ with $P(x, y_1, \ldots, y_m) = 0$, where P is a polynomial. Suppose we were to possess an algorithm to determine if an arbitrary Diophantine equation has natural number solutions. Given x, we would then be able to use the algorithm to determine whether $P(x, y_1, \ldots, y_m)$ has a solution in $y_1, \ldots, y_m$. That is, we would have an algorithm to decide whether or not x is in U. This implies U is decidable, a contradiction. Hence no such algorithm exists! □

While satisfying, the above argument relies on our construction in Theorem 4.2 of the set U, which made use of the halting problem. It is possible to describe such a set directly in terms of Diophantine sets. In the remainder of this section, we do this to give a solution to Hilbert's tenth problem that does not make reference to the halting problem.

Our first step is to enumerate the Diophantine sets. Since Diophantine sets are described by polynomials, we begin by enumerating all polynomials. Let us write our polynomials with variables $x_0, x_1, \ldots$. Inductively, we define

$$\begin{aligned} P_0 &= 1, \\ P_{3i+1} &= x_i, \\ P_{3i+2} &= P_{L(i)} + P_{R(i)}, \text{ and} \\ P_{3i+3} &= P_{L(i)} P_{R(i)}. \end{aligned}$$

Every three steps, this enumeration adds a new variable and lists the sum and product of two previously listed polynomials. It does this in such a way that every variable is eventually listed, and the sum and product of every pair of listed polynomials are eventually listed. Since all polynomials with natural number coefficients can be built by taking sums and products of finitely many variables and 1, this

enumeration will list all such polynomials. We note that our enumeration is redundant, as any given polynomial will appear infinitely often.[3] Now, every Diophantine equation can be written as $P_i = P_j$ for some pair i, j, and we can rewrite this using a single index with the functions L and R, namely $P_{L(n)} = P_{R(n)}$ for some value of n. Thus all Diophantine sets of natural numbers can be enumerated explicitly as D_n for D_n equal to

$$\left\{x_0 : (\exists x_1, \ldots, x_n) \left(P_{L(n)}(x_0, x_1, \ldots, x_n) = P_{R(n)}(x_0, x_1, \ldots, x_n)\right)\right\}.$$

Since $L(n)$ and $R(n)$ are at most n, $P_{L(n)}$ and $P_{R(n)}$ can only depend on variables up to x_n. Note that since the Diophantine sets are the computably enumerable sets, D_n also gives an enumeration of the computably enumerable sets.

Using Cantor's method of diagonalization, we can use this enumeration of Diophantine sets to construct a set different from each D_n, and hence is not Diophantine.

Theorem 5.27. *There is a set V of natural numbers that is not Diophantine.*

Proof. We let $V = \{n : n \notin D_n\}$. Suppose V were Diophantine. We would then have $V = D_k$ for some k. Then

$$k \in D_k \iff k \in V \iff k \notin D_k,$$

the latter equivalence by the definition of V. This is a contradiction, and so V is not a Diophantine set. □

Note that since the Diophantine sets are the computably enumerable sets, V is an example of a set that is not computably enumerable. We previously gave an example of such a set in Theorem 4.2 by use of the halting problem.

We now show that the set of ordered pairs (n, x) with $x \in D_n$ is Diophantine. We call this the *universality theorem.*

[3] For example, since $P_0 = 1$, for any polynomial P_n we have $P_0 P_n = P_n$. Since the Cantor pairing function applied to $(0, n)$ is equal to $\frac{n(n+1)}{2}$, we have $L\left(\frac{n(n+1)}{2}\right) = 0$ and $R\left(\frac{n(n+1)}{2}\right) = n$. Thus $P_{\frac{3n(n+1)}{2}+3} = P_0 P_n = P_n$. This process may then be repeated.

Theorem 5.28 (Universality theorem)**.** *The set*

$$\{(n, x) : x \in D_n\}$$

is Diophantine

Proof. We show that $x \in D_n$ if and only if there exists u such that each of the following hold:

D1. $\beta(u, 0) = 1$,

D2. $\beta(u, 1) = x$,

D3. $(\forall i)_{\leq n} \left(\beta(u, 3i + 2) = \beta(u, L(i)) + \beta(u, R(i))\right)$,

D4. $(\forall i)_{\leq n} \left(\beta(u, 3i + 3) = \beta(u, L(i))\beta(u, R(i))\right)$,

D5. $\beta(u, L(n)) = \beta(u, R(n))$.

Since each of these are Diophantine, it will follow that $\{(n, x) : x \in D_n\}$ is Diophantine.

To show the forward direction, suppose $x \in D_n$. By definition of D_n, there exist $x_1, \ldots, x_n$ with

$$P_{L(n)}(x, x_1, \ldots, x_n) = P_{R(n)}(x, x_1, \ldots, x_n).$$

We can then use the β function to choose u with

$$\beta(u, i) = P_i(x, x_1, \ldots, x_n) \quad \text{for } i = 0, \ldots, 3n + 3.$$

By definition of the P_i, conditions D1 to D4 hold. Since

$$P_{L(n)}(x, x_1, \ldots, x_n) = P_{R(n)}(x, x_1, \ldots, x_n),$$

D5 holds, completing the proof of this direction.

To show the converse, suppose D1 to D5 hold for some n, x, and u. Let $x_1 = \beta(u, 4)$, $x_2 = \beta(u, 7)$, $\ldots$, $x_n = \beta(u, 3n + 1)$. Then, by construction of the P_i with D1 to D4, we have $\beta(u, i) = P_i(x, x_1, \ldots, x_n)$ for $i = 0, \ldots, 3n + 3$. By D5, it follows that

$$P_{L(n)}(x, x_1, \ldots, x_n) = P_{R(n)}(x, x_1, \ldots, x_n).$$

By definition of D_n, this implies $x \in D_n$, completing the proof. □

We are now ready to give an alternative proof of Theorem 5.26.

Alternative Proof of Theorem 5.26. By the universality theorem, we have $x \in D_n$ if and only if there exist $y_1, \ldots, y_m$ such that

$$P(x, n, y_1, \ldots, y_m) = 0.$$

Here P is some particular (fixed) polynomial.[4] Suppose now that we were to possess an algorithm to determine if an arbitrary Diophantine equation has natural number solutions. We could use this algorithm to test P for solutions. That is, we could use the algorithm to determine whether or not $x \in D_n$. This algorithm could then be used to list the elements of the non-Diophantine set V of Theorem 5.27: for each natural number n, determine if $n \in D_n$. If it is not, then we list n. This would imply that V is a computably enumerable set, which is a contradiction. Thus no such algorithm exists. □

We note that the negative solution to Hilbert's tenth problem implies that there is no one algorithm that can decide whether or not an arbitrary Diophantine equation has solutions. There are, of course, algorithms that work for certain classes of Diophantine equations.

To end this section, we briefly discuss the history of the solution to Hilbert's tenth problem. In his PhD thesis, Martin Davis showed that every computably enumerable set can be put in what is called the *Davis normal form*, where only existential quantifiers and one bounded universal quantifier are used.

Julia Robinson worked with Diophantine sets, attempting to show that the exponential function is Diophantine. Unable to do so, she considered the *exponential Diophantine sets*. A set S of ordered n-tuples is exponential Diophantine if there is a polynomial

$$P(x_1, \ldots, x_n, u_1, \ldots, u_m, v_1, \ldots, v_m, w_1, \ldots, w_m)$$

such that $(x_1, \ldots, x_n) \in S$ if and only if there exists u_i, v_i, and w_i with $P = 0$ and $u_i = v_i^{w_i}$ for each i. She showed that the binomial coefficients and factorial function are exponential Diophantine using more or less the same method that we have used here.

Davis and Hilary Putnam worked together, and came upon the idea of using Gödel's β function to deal with the bounded universal

[4] In fact, James P. Jones has explicitly written down several such polynomials; see [**Jo1**] and [**Jo2**].

quantifier. They showed that, under the assumption that there are arbitrarily long arithmetic progressions consisting entirely of prime numbers, every computably enumerable set is exponential Diophantine. However, their assumption was an open problem at the time, one that was not resolved until 2004 (published in 2008 in [**GT**]) by Ben Green (born 1977) and Terrence Tao (born 1975). Robinson was able to remove this assumption. In [**DPR**], Davis, Putnam, and Robinson proved that the computably enumerable sets are exponential Diophantine in a 1961 paper.

All that remained was to show that the exponential function is Diophantine. Given the consequences of this, some at the time felt this was unlikely. In 1970, Matiyasevič was able to show that the Fibonacci numbers are Diophantine, from which it follows that the exponential function is Diophantine. In this text, we have used solutions to the Brahmagupta–Pell equation in place of the Fibonacci numbers, but the methods are similar. In particular, it was Matiyasevič who first saw how to use results similar to our Lemmas 5.5 and 5.11.

Davis, Matiyasevič, Putnam, and Robinson all played crucial roles in the solution of the problem. Davis, Matiyasevič, and Robinson continued to work on the consequences of their solution to Hilbert's tenth problem, and some of their work is described in the next chapter.

Further Reading

Martin Davis's article *Hilbert's tenth problem is unsolvable* [**Da2**] is an excellent and approachable exposition of the solution to the tenth problem; indeed, it has influenced our presentation of the material here. Note that while we include 0 as a natural number here, Davis does not in his paper. Our presentation here uses some modifications given by Jones et al. in [**JSWW**].

Matiyasevič's book *Hilbert's Tenth Problem* [**Ma**] is another excellent source. The way in which he approaches the problem is somewhat different from our approach, and he gives much additional information that we do not cover.

If you used Hodel's *An Introduction to Mathematical Logic* [**Ho**] to learn more about the material from Sections 4.3 and 4.4, then note that the final chapter of his book covers Hilbert's tenth problem.

Smoryński's *Logical Number Theory I* [**Sm**] also gives a solution to Hilbert's tenth problem. This interesting book contains numerous results in the intersection of logic and number theory.

Exercises

5.1. Use the definition of a Diophantine set to show that each of the following sets of natural numbers are Diophantine. In particular, write down a polynomial $P(x, y_1, \ldots, y_m)$ such that there exist $y_1, \ldots, y_m$ with $P(x, y_1, \ldots, y_m) = 0$ if and only if x is an element of the given set.
   (a) The set of even numbers.
   (b) The multiples of 3.
   (c) Natural numbers that are not multiples of 3.
   (d) Natural numbers that belong to a Pythagorean triple.
   (e) The set of perfect squares.
   (f) The set of natural numbers that are not perfect squares. (*Hint*: Recall that when the natural number d is a nonzero perfect square, the Brahmagupta–Pell equation $x^2 - dy^2 = 1$ has only the trivial solution $(x, y) = (1, 0)$, and when d is not a perfect square, there exists a nontrivial solution.)

5.2. Let A and B be Diophantine sets of n-tuples. Show that $A \cap B$ and $A \cup B$ are Diophantine.

5.3. For $r \in \mathbb{R}$, let $\lfloor r \rfloor$ denote the greatest integer less than or equal to r. Prove that for natural numbers x, y, and z, $z = \lfloor x/y \rfloor$ if and only if

$$yz \leq x < y(z + 1).$$

Use this to prove that the function $f(x, y) = \lfloor x/y \rfloor$ is Diophantine.

5.4. Show that $L(z) \leq z$ and $R(z) \leq z$.

5.5. Find a value of u such that $\beta(u,i)$ of Theorem 5.2 satisfies $\beta(u,0)=2$, $\beta(u,1)=3$, and $\beta(u,2)=5$.

5.6. Fix a natural number b. Prove that

$$\lim_{a\to\infty} \frac{a^b}{\binom{a}{b}} = b!.$$

Although a^b and $\binom{a}{b}$ are Diophantine, the above limit does not immediately imply that $b!$ is Diophantine. In Theorem 5.18, we showed that taking a large enough (we took $a=(2b)^b$) brings $a^b/\binom{a}{b}$ within distance 1 of $b!$, which allowed us to show that $b!$ is Diophantine.

5.7. We say that n is a lower twin prime if n and $n+2$ are both prime. Make use of either Wilson's theorem or the bounded universal quantifier to show that the set of lower twin primes is Diophantine. Deduce that there is a polynomial whose nonnegative range is equal to the set of lower twin primes.

Chapter 6

Applications of Hilbert's Tenth Problem

6.1. Related Problems

In this section, we discuss a few corollaries and relatively immediate applications of the negative solution to Hilbert's tenth problem.

If we let N_P be the number of solutions to a Diophantine equation $P = 0$, then Hilbert's tenth problem asks for an algorithm to determine if $N_P = 0$. Given $k \in \mathbb{N} \cup \{\aleph_0\}$ with $k \neq 0$, is there an algorithm that will determine if $N_P = k$? We show now that the answer to this question is no.[1]

Theorem 6.1. *Let $k \in \mathbb{N} \cup \{\aleph_0\}$ with $k \neq 0$. Then is no algorithm that will determine if an arbitrary Diophantine equation has exactly k solutions.*

Proof. We first consider the case $k = \aleph_0$. Let y be a variable not appearing in P, and let $Q = (y+1)P$. We show $N_Q = \aleph_0$ if and only if $N_P > 0$. Suppose $N_Q = \aleph_0$ so that $(y+1)P = 0$ has infinitely many solutions. Since $y+1 \neq 0$, $P = 0$ must have at least one solution, and so $N_P > 0$. Conversely, suppose $N_P > 0$. Then there is a solution

[1] This result is due to Martin Davis and can be found in [**Da3**]. In fact, he has shown more: let S be a nonempty proper subset of $\{\aleph_0, 0, 1, \ldots\}$. Then there is no algorithm to determine whether or not $N_P \in S$.

to $P = 0$, which yields a solution to $(y+1)P = 0$ for $y = 0, 1, 2, \ldots$, and hence $N_Q = \aleph_0$. Thus $N_Q \neq \aleph_0$ if and only if $N_P = 0$. Suppose we were to possess an algorithm to test whether or not a Diophantine equation has infinitely many solutions. Given a Diophantine equation P, the algorithm would allow us to determine if $N_Q \neq \aleph_0$, and hence if $N_P = 0$. This contradicts the negative solution to Hilbert's tenth problem.

We now consider the case $k \in \mathbb{N}$. By $n-1$ applications of the inverse of Cantor's pairing function, we may computably enumerate the n-tuples of natural numbers. Given a nonzero Diophantine equation $P = P(x_1, \ldots, x_n)$, we define Diophantine equations P_k inductively by $P_0 = P$ and

$$P_{k+1} = P_k \cdot \left((x_1 - a_1)^2 + \cdots + (x_n - a_n)^2\right),$$

where $(a_1, \ldots, a_n)$ is the first n-tuple that is not a solution to P_k. Then $P_{k+1} = 0$ if and only if $P_k = 0$ or $(x_1, \ldots, x_n) = (a_1, \ldots, a_n)$, and it cannot be the case that both of these disjuncts are true. Thus $N_{P_{k+1}} = N_{P_k} + 1$, and so $N_{P_k} = N_P + k$. Therefore, $N_P = 0$ if and only if $N_{P_k} = k$. Suppose we were to possess an algorithm to test whether or not a Diophantine equation has exactly k solutions. Given a Diophantine equation P, the algorithm would allow us to determine if $N_{P_k} = k$, and hence if $N_P = 0$. This contradicts the negative solution to Hilbert's tenth problem. □

In Chapter 5, we saw that conjunction, disjunction, the existential quantifier, and the bounded universal quantifier may be used on Diophantine expressions to yield another Diophantine expression. We can now show that use of any of the remaining standard logical connectives or of the (unbounded) universal quantifier on Diophantine expressions may yield a non-Diophantine expression. In Section 5.7, we gave an enumeration D_n of all Diophantine sets of natural numbers. The universality theorem (Theorem 5.28) established that the set

$$\{(n, x) : x \in D_n\}$$

is Diophantine. Thus there is a polynomial P such that $x \in D_n$ if and only if there exist $y_1, \ldots, y_m$ with $P(x, n, y_1, \ldots, y_m) = 0$. Finally,

in Theorem 5.27 the set $V = \{n : n \notin D_n\}$ was shown to be not Diophantine. Thus the following are equivalent:

(a) $n \in V$.

(b) It is not the case that $(\exists y_1, \ldots, y_m)(P(n, n, y_1, \ldots, y_m) = 0)$.

(c) If $(\exists y_1, \ldots, y_m)(P(n, n, y_1, \ldots, y_m) = 0)$, then $0 = 1$.

(d) For all $y_1, \ldots, y_m$,

$$P(n, n, y_1, \ldots, y_m) < 0 \quad \text{or} \quad P(n, n, y_1, \ldots, y_m) > 0.$$

Now, (b) is negation applied to a Diophantine expression, (c) is the implication of two Diophantine expressions, and (d) is the universal quantification of a Diophantine expression. Thus the negation, implication, and universal quantification of Diophantine expressions may yield an expression that is not Diophantine.

Given a Diophantine set of natural numbers, we may ask for the *smallest* polynomial that can be used in its Diophantine description. We define the *degree* of the Diophantine set

$$S = \{x : (\exists y_1, \ldots, y_m)(P(x, y_1, \ldots, y_m) = 0)\}$$

to be the least degree of a polynomial P that may be used in a Diophantine description of S. We define the *dimension* of S to be the least value of m that may be used in such a P.

We may reduce the degree of an equation $P(x_1, \ldots, x_n) = 0$ as follows. We introduce new equations $z_i = x_j x_k$ and $z_i = x_j^2$ with new variables z_i. By applying repeated substitution with these new equations, the degree of P may be brought down to 2. Then the equation found by summing the squares of the additional equations with the square of the reduced P will be equal to 0 if and only if $P = 0$, and it will be of degree 4. As an example, consider the equation

$$x_1^3 x_2 x_3^4 - 2x_1^2 = 0.$$

We set $z_1 = x_1^2$, $z_2 = x_1 z_1$ so that $z_2 = x_1^3$, $z_3 = x_3^2$, $z_4 = z_3^2$ so that $z_4 = x_3^4$ and $z_5 = x_2 z_4$. Then $x_1^3 x_2 x_3^4 - 2x_1^2 = z_2 z_5 - 2z_1$. Thus the original equation is equal to 0 if and only if

$$(z_1 - x_1^2)^2 + (z_2 - x_1 z_1)^2 + (z_3 - x_3^2)^2 + (z_4 - z_3^2)^2 + (z_5 - x_2 z_4)^2 + (z_2 z_5 - 2z_1)^2 = 0.$$

Note that our method for decreasing the degree will increase the number of variables. We have shown the following theorem.

Theorem 6.2. *Any Diophantine set of natural numbers may be described with a polynomial of degree at most* 4.

There is an algorithm that will determine if an arbitrary Diophantine equation of degree 2 has natural number solutions, given in 1972 by Carl Ludwig Siegel (1896–1981) in [**Sie**]. Hence, not every Diophantine set may be described with a degree 2 polynomial, since this would yield a positive solution to Hilbert's tenth problem. Whether or not every Diophantine set may be described with a degree 3 polynomial remains an open question. It is related to the theory of elliptic curves.

It turns out that the dimension of a Diophantine set of natural numbers also has an upper bound.

Theorem 6.3. *There exists a natural number m such that any Diophantine set of natural numbers may be described with a polynomial with at most $m + 1$ variables.*

Proof. Recall that in Section 5.7, we gave an enumeration D_n of the Diophantine sets of natural numbers. The universality theorem (Theorem 5.28) yields a *universal polynomial* P such that x is an element of D_n if and only if there exist $y_1, \ldots, y_m$ with $P(x, n, y_1, \ldots, y_m) = 0$. Given a Diophantine set S of natural numbers, there is some n for which $S = D_n$. That is, we have

$$S = \{x : (\exists y_1, \ldots, y_m)(P(x, n, y_1, \ldots, y_m) = 0)\}$$

for some particular value of n. Thus, the value of m in such a polynomial P gives an upper bound on the dimension of any Diophantine set of natural numbers. □

One could use the results of Chapter 5 to write down such a polynomial P, although it would be tedious and the result would be far from optimal.[2] In fact, it has been shown that m can be brought

[2]Several universal polynomials have been explicitly written down by James P. Jones in [**Jo1**] and [**Jo2**].

as low as 9.[3] It is possible that an even lower value would suffice, although this is an open question.

The existence of a universal polynomial P is truly surprising. By merely changing the value of n, the polynomial P will define any Diophantine set. Since the Diophantine sets are one and the same as the computably enumerable sets, P defines all computably enumerable sets. The computably enumerable sets are those for which we have an algorithm to list the elements. It is surprising that a single polynomial P can contain enough information to yield every possible set of natural numbers for which we have an element listing algorithm![4]

6.2. A Prime Representing Polynomial

In Theorem 5.25, we saw that the Diophantine sets are one and the same as the computably enumerable sets. Since we have an algorithm to list the prime numbers, they form a Diophantine set. In Theorem 5.1, we saw that a set is Diophantine if and only if it is the nonnegative range of a polynomial. This implies the existence of a polynomial whose nonnegative range consists of the set of prime numbers (the negative numbers in the range of the polynomial may or may not be prime). In fact, the methods we gave in Chapter 5 were constructive, allowing such a polynomial to be written down. In general, writing out these polynomials can be tedious, especially if bounded universal quantifiers must be used. However, Wilson's theorem (Theorem 3.11) allowed us to give a Diophantine definition of the prime numbers without any bounded universal quantifiers. In this section, we use Wilson's theorem to give an explicit Diophantine definition of the

[3]This was proved by Matiyasevič in 1975, partially using methods he developed with Robinson. The result remained unpublished for seven years until Jones gave a proof, with permission from Matiyasevič, in [**Jo2**].

[4]This may not have been surprising to Turing. In the same paper in which he first introduced Turing machines, Turing proved the existence of a *universal Turing machine*: a Turing machine $U(m, n)$ that, given n, will run the Turing machine T_n on input m. Here T_n is the nth machine in the enumeration of the Turing machines given in Theorem 4.1.

prime numbers, and then write down an explicit *prime representing* polynomial.[5]

Recall that Wilson's theorem states that a natural number $n > 1$ is prime if and only if

$$(n-1)! \equiv -1 \pmod{n}.$$

Letting $n = m + 2$ for a natural number m, we have $m + 2$ is prime if and only if

$$(m+1)! \equiv -1 \pmod{m+2}.$$

This occurs if and only if

$$\begin{aligned}(m+2) \mid ((m+1)!+1) &\\ &\iff (\exists \kappa \in \mathbb{N})((m+1)!+1 = (\kappa+1)(m+2)) \\ &\iff (\exists \kappa \in \mathbb{N})((m+1)! = \kappa m + 2\kappa + m + 1).\end{aligned}$$

By Theorem 5.18, $\beta = (m+1)!$ if and only if there exist h, n, q, z, γ, δ, and ϵ satisfying

F1. $z = 2(m+1)$,
F2. $h = n + 1$,
F3. $h = z^{m+1}$,
F4. $\gamma = \binom{n+1}{m+1}$,
F5. $q = h^{m+1}$,
F6. $\beta\gamma + \delta = q$,
F7. $q + \epsilon + 1 = (\beta+1)\gamma$.

These equations have been slightly modified from those of Theorem 5.18; inequalities have been replaced with their Diophantine equivalent and some variables have been renamed. Note that the value of a in Theorem 5.18 is at least 2, and hence not 0, allowing us to replace it with $n + 1$.

By Theorem 5.17, $\gamma = \binom{n+1}{m+1}$ if and only if there exist f, g, x, y, α, ϕ, ρ, and σ satisfying

C1. $x = 3^{n+1}$,

[5]Such a polynomial was first explicitly given in 1976 by Jones, Sato, Wada, and Wiens in [**JSWW**], and we follow their method here. Due to differences in the presentation of material, our polynomial will be slightly longer than the one they gave.

C2a. $y = x + 1$,
C2b. $g = y^{n+1}$,
C2c. $f = x^{m+1}$,
C2d. $g = \phi x f + \gamma f + \rho$,
C3. $\gamma + \alpha + 1 = x$,
C4. $\rho + \sigma + 1 = f$.

Equation C2 of Theorem 5.17 has been split into multiple equations, inequalities have been replaced with their Diophantine equivalent, and variables have been renamed. In Theorem 5.17, the only requirement on u (here called x) is that it is greater than 2^{n+1}. We take $x = 3^{n+1}$ for convenience.

Thus $m + 2$ is prime if and only if there exist $f, g, h, n, q, x, y, z, \alpha, \beta, \gamma, \delta, \epsilon, \kappa, \phi, \rho$, and σ satisfying $\beta = \kappa m + 2\kappa + m + 1$ and equations F1, F2, F3, F5, F6, F7, C1, C2a, C2b, C2c, C2d, C3, and C4. All but F3, F5, C1, C2b, and C2c are Diophantine equations. Thus we must replace $h = z^{m+1}$, $q = h^{m+1}$, $x = 3^{n+1}$, $g = y^{n+1}$, and $f = x^{m+1}$ with their Diophantine equivalents. We could do this by using Theorem 5.15 five times, which will result in 35 more Diophantine equations and 45 new variables. However, it is possible to modify Theorem 5.15 to encode multiple exponentials at less of a cost. In the theorem below, equations X1 to X7 are essentially E1 to E7 of Theorem 5.15, although we have modified E6. They encode $f = x^{n+1}$. Equations X8 to X10 below allow us to introduce a second exponential $g = y^{m+1}$ by re-using some of the work of equations X1 to X7, although we will also require that $m < n$ (as is the case here). These three equations use four new variables. Finally, each of equations X11 to X13 allow us to add an additional exponential that uses exponents $n + 1$ or $m + 1$. Each of these equations introduces one new variable.

Theorem 6.4. *Given numbers f, g, h, m, n, q, x, y, and z with $m < n$ and $h, x, y, z \geq 1$, the system*

X1. $v^2 - (a^2 - 1)w^2 = 1$,
X2. $u^2 - 16(a^2 - 1)r^2w^4 = 1$,
X3. $(v + cu)^2 - ((a + u^2(u^2 - a))^2 - 1)(n + 1 + 4dw)^2 = 1$,
X4. $w = n + s + p + 1$,

X5. $e^3(e+2)(a+1)^2+1=\ell^2$,

X6. $e=x+y+m+n+f+g+h+q+z+2$,

X7. $v=f+w(a-x)+i(2ax-x^2-1)$,

X8. $t^2-(a^2-1)s^2=1$,

X9. $s=m+1+k(a-1)$,

X10. $t=g+s(a-y)+j(2ay-y^2-1)$,

X11. $v=x+w(a-3)+\lambda(6a-10)$,

X12. $t=h+s(a-z)+\chi(2az-z^2-1)$,

X13. $t=q+s(a-h)+\omega(2ah-h^2-1)$

has a solution in the remaining variables if and only if $f=x^{n+1}$, $g=y^{m+1}$, $h=z^{m+1}$, $q=h^{m+1}$, *and* $x=3^{n+1}$.

Proof. We show the forward direction first. Suppose equations X1 to X13 are satisfied. Equation X6 implies $e\geq 2$, and so Lemma 5.14 with X5 implies $e-1+e^{e-2}\leq a$. With X6, this implies

$$x+y+m+n+f+g+h+q+z+1 \\ +\big(x+y+m+n+f+g+h+q+z+2\big)^{x+y+m+n+f+g+h+q+z}\leq a.$$

Using this, we deduce that the following quantities are strictly less than a: f, g, h, q, x, x^{n+1}, y^{m+1}, z^{m+1}, h^{m+1}, and 3^{n+1}. Also, we have $a>2$, $n<a-1$, $m+1<a-1$, and $y\leq a-1$. As in the proof of Theorem 5.15, equations X1 to X4 imply $w=y_{n+1}(a)$ and $v=x_{n+1}(a)$. The conclusion that $f=x^{n+1}$ may be drawn using Lemma 5.8 with equation X7, as in the proof of Theorem 5.15.

Equation X8 implies $s=y_{m'}(a)$ and $t=x_{m'}(a)$ for some m'. Equation X4 yields $s<w$, and so $m'<n+1$. Since $n<a-1$, we have $m'\leq n<a-1$. Equation X9 implies

$$s\equiv m+1 \pmod{a-1},$$

and Lemma 5.7 implies

$$s\equiv m' \pmod{a-1}.$$

Since $m'\equiv m+1 \pmod{a-1}$ with $m'<a-1$ and $m+1<a-1$, we have $m'=m+1$ and so

$$s=y_{m+1}(a) \quad \text{and} \quad t=x_{m+1}(a).$$

Then Lemma 5.8 implies

$$t \equiv y^{m+1} + s(a-y) \pmod{2ay - y^2 - 1},$$

while equation X10 implies

$$t \equiv g + s(a-y) \pmod{2ay - y^2 - 1}.$$

Hence,

$$g \equiv y^{m+1} \pmod{2ay - y^2 - 1}.$$

Since $y \leq a-2$, we have

$$y^2 + 1 \leq y(a-2) + 1 = ay + 1 - 2y.$$

Now $y \geq 1$ yields both $y \leq 2y - 1$ and $1 - 2y < 0$. Thus $y^2 + 1 < a(2y-1)$, and so

$$a < 2ay - y^2 - 1.$$

Since $g < a$ and $y^{m+1} < a$, we have both g and y^{m+1} less than $2ay - y^2 - 1$. Since they are also congruent modulo $2ay - y^2 - 1$, it follows that $g = y^{m+1}$, as required.

Now Lemma 5.8 yields

$$v \equiv 3^{n+1} + w(a-3) \pmod{6a-10},$$

while equation X11 yields

$$v \equiv x + w(a-3) \pmod{6a-10}.$$

Thus,

$$x \equiv 3^{n+1} \pmod{6a-10}.$$

Since $x < a < 6a - 10$ and $3^{n+1} < a < 6a - 10$, we have $x = 3^{n+1}$. Similarly, Lemma 5.8 yields

$$t \equiv z^{m+1} + s(a-z) \pmod{2az - z^2 - 1}$$

and

$$t \equiv h^{m+1} + s(a-h) \pmod{2ah - h^2 - 1},$$

while equation X12 yields

$$t \equiv h + s(a-z) \pmod{2az - z^2 - 1}$$

and X13 yields

$$t \equiv q + s(a-h) \pmod{2ah - h^2 - 1}.$$

Thus

$$h \equiv z^{m+1} \pmod{2az - z^2 - 1}$$

and

$$q \equiv h^{m+1} \pmod{2ah - h^2 - 1}.$$

Since $h < a < 2az - z^2 - 1$, $z^{m+1} < a < 2az - z^2 - 1$, $q < a < 2ah - h^2 - 1$, and $h^{m+1} < a < 2ah - h^2 - 1$, we have $h = z^{m+1}$ and $q = h^{m+1}$.

We now show the reverse direction. Suppose $f = x^{n+1}$, $g = y^{m+1}$, $h = z^{m+1}$, $q = h^{m+1}$, and $x = 3^{n+1}$ for $m < n$ and $h, x, y, z \geq 1$. Set

$$e = x + y + m + n + f + g + h + q + z + 2$$

to satisfy X6. Equations X1, X2, X3, X5, and X7 may be satisfied as in the proof of Theorem 5.15, with $n < w$. Set $s = y_{m+1}(a)$ and $t = x_{m+1}(a)$ so that X8 is satisfied. Since $y_k(a)$ is increasing with k and we have $m + 1 \leq n$, it follows that $y_{m+1}(a) \leq y_n(a)$. One may show with induction and Theorem 5.3 that $n + y_n(a) < y_{n+1}(a)$. Thus,

$$\begin{aligned} n + s &= n + y_{m+1}(a) \\ &\leq n + y_n(a) \\ &< y_{n+1}(a) \\ &= w, \end{aligned}$$

and so p may be found to satisfy X4. Lemma 5.7 implies $s \equiv m + 1 \pmod{a - 1}$. Since (5.5) of Theorem 5.4 yields $m + 1 \leq y_{m+1}(a) = s$, we may choose k so that X9 is satisfied. Lemma 5.8 implies

$$\begin{aligned} t &\equiv g + s(a - y) \pmod{2ay - y^2 - 1}, \\ v &\equiv x + s(a - 3) \pmod{6a - 10}, \\ t &\equiv h + s(a - z) \pmod{2az - z^2 - 1}, \text{ and} \\ t &\equiv q + s(a - h) \pmod{2ah - h^2 - 1}, \end{aligned}$$

with each right-hand side less than or equal to the left-hand side, This allows us to choose j, λ, χ, and ω to satisfy equations X10, X11, X12, and X13. □

To summarize what we have so far, $m+2$ is prime if and only if $\beta = \kappa m + 2\kappa + m + 1$, and equations F1, F2, F6, F7, C2a, C2d, C3, C4, and X1 to X13 have a solution. This is a system of 22 equations in 36 variables (including m). We can eliminate some of the variables with substitution, provided the quantity we are substituting is always positive (otherwise, we may be losing information). We use $\beta = \kappa m+2\kappa+m+1$ and equations F1, F2, F6, C2a, C2d, C3, C4, X4, and X9 to eliminate variables $\beta, z, h, q, y, g, x, f, w$, and s. Equation F7 then becomes $\gamma = \delta + \epsilon + 1$, allowing us to further eliminate γ. This leaves us with 11 equations in 25 variables. We have $m+2$ prime if and only if the following system has a solution:

D1. $v^2 - (a^2-1)(n+m+p+k(a-1)+2)^2 = 1,$

D2. $u^2 - 16r^2(a^2-1)(n+m+p+k(a-1)+2)^4 = 1,$

D3. $(v+cu)^2 - ((a+u^2(u^2-a))^2-1)$
$\times(n+1+4d(n+m+p+k(a-1)+2))^2 = 1,$

D4. $e^3(e+2)(a+1)^2 + 1 = \ell^2,$

D5. $e = 3m+2n+2\alpha+\delta+2\rho+\sigma+\phi(\rho+\sigma+1)(\alpha+1)$
$+(\kappa m+2\kappa+m+(\phi+1)(\rho+\sigma+1)+3)(\delta+\epsilon+1)+9,$

D6. $v = \rho+\sigma+1+(n+m+p+k(a-1)+2)(a-\delta-\epsilon-\alpha-2)$
$+i(2a(\delta+\epsilon+\alpha+2)-(\delta+\epsilon+\alpha+2)^2-1),$

D7. $t^2 - (a^2-1)(m+1+k(a-1))^2 = 1,$

D8. $t = (\rho+\sigma+1)(\phi(\alpha+1)+(\phi+1)(\delta+\epsilon+1))+\rho$
$+(m+1+k(a-1))(a-\delta-\epsilon-\alpha-3)$
$+j(2a(\delta+\epsilon+\alpha+3)-(\delta+\epsilon+\alpha+3)^2-1),$

D9. $v = \delta+\epsilon+\alpha+2+(n+m+p+k(a-1)+2)(a-3)+\lambda(6a-10),$

D10. $t = n+1+(m+1+k(a-1))(a-2m-2)$
$+\chi(4a(m+1)-4(m+1)^2-1),$

D11. $t = (\kappa m+2\kappa+m+1)(\delta+\epsilon+1)+\delta$
$+(m+1+k(a-1))(a-n-1)+\omega(2a(n+1)-(n+1)^2-1).$

Some of these equations have been simplified after making the substitutions.

We can rearrange each equation so that all terms are on one side. For the time being, let us write the resulting equations as $A = 0$, $B = 0, \ldots, K = 0$. Consider the polynomial

$$(m+2)(1 - A^2 - B^2 - \cdots - K^2).$$

If $m+2$ is prime, then $A = B = \cdots = K = 0$, and so $m+2$ is in the (positive) range of this polynomial. On the other hand, if ψ is in the positive range of the polynomial, then we have

$$\psi = (m+2)(1 - A^2 - B^2 - \cdots - K^2).$$

For the right-hand side to be positive, we must have

$$A = B = \cdots = K = 0.$$

This means $m+2$ must be prime and yields $\psi = m+2$. Thus

$$(m+2)(1 - A^2 - B^2 - \cdots - K^2)$$

is a polynomial whose positive range is equal to the set of prime numbers.

Using our equations above, we may now write down our prime representing polynomial. Since we have eliminated some variables, we may replace the greek letter variables with now unused roman letters. We replace

$$\alpha, \delta, \epsilon, \kappa, \lambda, \rho, \sigma, \phi, \chi, \text{ and } \omega$$

with

$$b, f, g, h, q, s, w, x, y, \text{ and } z,$$

respectively.

Theorem 6.5. *The nonnegative range of*

$$
\begin{aligned}
(m+2)\Big[1 &- \big(v^2-(a^2-1)(n+m+p+k(a-1)+2)^2-1\big)^2\\
&- \big(u^2-16r^2(a^2-1)(n+m+p+k(a-1)+2)^4-1\big)^2\\
&- \big((v+cu)^2-((a+u^2(u^2-a))^2-1)\\
&\qquad\cdot(n+1+4d(n+m+p+k(a-1)+2))^2-1\big)^2\\
&- \big(e^3(e+2)(a+1)^2+1-\ell^2\big)^2\\
&- \big(3m+2n+2b+f+2s+w+x(s+w+1)(b+1)\\
&\qquad+(hm+2h+m+(x+1)(s+w+1)+3)\\
&\qquad\qquad\cdot(f+g+1)+9-e\big)^2\\
&- \big(s+w+1+(n+m+p+k(a-1)+2)\\
&\qquad\cdot(a-f-g-b-2)\\
&\qquad+i(2a(f+g+b+2)-(f+g+b+2)^2-1)-v\big)^2\\
&- \big(t^2-(a^2-1)(m+1+k(a-1))^2-1\big)^2\\
&- \big((s+w+1)(x(b+1)+(x+1)(f+g+1))+s\\
&\qquad+(m+1+k(a-1))(a-f-g-b-3)\\
&\qquad+j(2a(f+g+b+3)-(f+g+b+3)^2-1)-t\big)^2\\
&- \big(f+g+b+2+(n+m+p+k(a-1)+2)\\
&\qquad\qquad\cdot(a-3)+q(6a-10)-v\big)^2\\
&- \big(n+1+(m+1+k(a-1))(a-2m-2)\\
&\qquad+y(4a(m+1)-4(m+1)^2-1)-t\big)^2\\
&- \big((hm+2h+m+1)(f+g+1)+f+(m+1+k(a-1))\\
&\qquad\cdot(a-n-1)+z(2a(n+1)-(n+1)^2-1)-t\big)^2\Big]
\end{aligned}
$$

is equal to the set of prime numbers.

This polynomial has 25 natural number variables, comprising all letters except for o.

One might try to use this polynomial to discover new primes, but unfortunately this is highly impractical. Since our original equations

C2b, C2a, C1, F2, F3, and F1 are $g = y^{n+1}$, $y = x + 1$, $x = 3^{n+1}$, $h = n + 1$, $h = z^{m+1}$, and $z = 2(m + 1)$, it follows that

$$g = \left(3^{(2(m+1))^{m+1}} + 1\right)^{(2(m+1))^{m+1}} \approx 3^{(2m+2)^{2m+2}}.$$

Along the way, g was replaced with a degree 3 polynomial in five of the greek letter variables, and so at least one of these variables must grow roughly as fast as the cube root of g. The size of g grows very fast with m. We could look for primes by running through all possible values of the variables of the polynomial, checking if the polynomial is positive, and noting the prime $m + 2$ if it is. Doing this, we would discover that 2 is prime once we had $g = (3^2 + 1)^2 = 100$ (and so at least one of the variables is around $\sqrt[3]{100} \approx 4.6$; not bad). We would discover that 3 is prime when $g = (3^{16} + 1)^{16}$. The cube root of this number has 41 digits. The next smallest prime, 5, would be discovered when $g \approx 3^{8^8} = 3^{16777216}$, a number whose cube root has 2668256 digits! The prime 7 requires a value of g whose cube root has over a trillion *digits*![6] For most values of its variables, the polynomial will be negative, and the extreme size of some of the variables needed to yield primes makes it impractical for finding primes.

6.3. Goldbach's Conjecture and the Riemann Hypothesis

The methods of Chapter 5 will allow us to show that many important open problems are equivalent to the unsolvability of a Diophantine equation. Furthermore, the constructive nature of the negative solution to Hilbert's tenth problem allows such polynomials to be written out explicitly, given enough time and patience.

To begin, we discuss *Goldbach's conjecture*. In 1742, Christian Goldbach (1690–1764), a mathematician best known today for his correspondence with some of the most prominent mathematicians of

[6] If you were to take the cube root of the value of g that yields the prime 11 and print it in 10-point font on rather thin (0.05 mm thick) letter-sized paper, the stack of paper would be higher than the distance travelled when making two return trips to Alpha Centauri! The cube root of the value of g that would yield the prime 17 would require a stack of paper 2.7 trillion times larger than the width of the *universe*!!

his day, proposed a conjecture to Euler:

> Every even natural number greater than 2 can be written as the sum of two primes.

Euler replied that he was certain this was true, but could not prove it. The conjecture remains open to this day. With the help of computers, it has been verified to hold for numbers up to 4×10^{18}. Several results have brought us close to the conjecture. In 1973, Chen Jingrun (1933–1996) showed that every sufficiently large even natural number can be written as either the sum of two primes or the sum of a prime and the product of two primes. Chen's proof uses sophisticated sieve theory and represents a *tour de force* of 20th-century analytic number theory. Earlier, in 1937, I. M. Vinogradov (1891–1983) applied the circle method, originally discovered by Srinivasa Ramanujan (1887–1920),[7] to show that every sufficiently large odd number can be written as the sum of three primes. His method was ineffective in that it supplied no explicit lower bound. Several mathematicians have devoted themselves to modifying Vinogradov's approach so as to make it effective. Recently, in 2013, building on earlier work by Olivier Ramaré and others, Harald Helfgott (born 1977) proved that every odd number greater than 5 can be written as the sum of three primes. This is known as the *ternary Goldbach problem*. It would be an easy consequence of the Goldbach conjecture, if the conjecture were proven to be true.

We now link Goldbach's conjecture with Diophantine equations.

Theorem 6.6. *There is a Diophantine equation that has no solutions if and only if Goldbach's conjecture is true.*

Proof. An even natural number greater than 2 can be written in the form $2a + 4$. As z runs from 0 to a,

$$(z + 2) + (2a + 2 - z)$$

runs over all possible ways of writing $2a+4$ as the sum of two natural numbers greater than or equal to 2. Thus, $2a+4$ is a counterexample

[7] The circle method was developed by G. H. Hardy (1877–1947) and Ramanujan in their study of partitions. Later, Hardy and J. E. Littlewood (1885–1977) applied it to study Waring's problem and other additive questions. Since then, it has been called the Hardly–Littlewood method.

to Goldbach's conjecture if and only if

$$(\forall z)_{\le a}(\exists x, y)\,(z+2=(x+2)(y+2) \text{ or } 2a+2-z=(x+2)(y+2))\,,$$

which is equivalent to

$$(\forall z)_{\le a}(\exists x, y)\big((z+2-(x+2)(y+2)) \cdot (2a+2-z-(x+2)(y+2))=0\big). \tag{6.1}$$

Applying the existential quantifier and the bounded universal quantifier to a Diophantine expression yields another Diophantine expression. Thus, there exists a polynomial $P(a, y_1, \ldots, y_m)$ such that $2a+4$ is a counterexample to Goldbach's conjecture if and only if there exist $y_1, \ldots, y_m$ with $P(a, y_1, \ldots, y_m)=0$. Hence, Goldbach's conjecture holds if and only if the Diophantine equation $P(a, y_1, \ldots, y_m) = 0$ has no solutions. □

Although it would be tedious, one can actually write down the above Diophantine equation. Using Theorem 5.21, the bounded universal quantifier can be removed from (6.1). Then earlier theorems can be used to replace the product functions, factorials, binomial coefficients, and exponential equations with Diophantine expressions. Recall that the resulting Diophantine equation $P = 0$ has natural number variables. If instead we desire a Diophantine equation that has no integer solutions if and only if Goldbach's conjecture is true, we can replace each natural number variable in P with the sum of the squares of four new integer variables.

The procedure that we have just applied to Goldbach's conjecture is actually quite general. We say that a property $\mathrm{Prop}(n)$ of the natural numbers is *decidable* if the set $S = \{n : \mathrm{Prop}(n) \text{ is true}\}$ is a decidable set. That is, a property is decidable if we have an algorithm that can be applied to any natural number to determine whether or not the property holds for that number. Note that this may tell us nothing about whether the property holds for *all* natural numbers, as we may only apply the algorithm to one n at a time.

Theorem 6.7. *Let* $\mathrm{Prop}(n)$ *be a decidable property of the natural numbers. There exists a Diophantine equation* $P(n, y_1, \ldots, y_m) = 0$ *such that* $\mathrm{Prop}(n)$ *holds for all* n *if and only if* $P = 0$ *has no solutions.*

Proof. Since Prop is decidable, we may run through the natural numbers n, determine whether Prop(n) holds for each, and list the n for which it does not hold. Thus, the set

$$\{n : \text{Prop}(n) \text{ is false}\}$$

is computably enumerable and, hence, Diophantine. This implies the existence of a polynomial $P(n, y_1, \ldots, y_m)$ such that Prop(n) is false if and only if there exist $y_1, \ldots, y_m$ for which $P(n, y_1, \ldots, y_m) = 0$. Finally, Prop($n$) is true for all n if and only if there does not exist an n for which Prop(n) is false. That is, if there do not exist $n, y_1, \ldots, y_m$ for which $P(n, y_1, \ldots, y_m) = 0$. □

The above theorem shows that any statement of the form "for all natural numbers, a decidable property holds" is equivalent to the unsolvability of a particular Diophantine equation. Since we may list the n for which a decidable property does not hold, these are statements for which there exists a process that will find a counterexample, if a counterexample exists. There are many important theorems and open problems that can be put in this form. Thus they are each equivalent to the unsolvability of a particular Diophantine equation! The problem of determining if a Diophantine equation has solutions is quite deep indeed. This makes the negative solution to Hilbert's tenth problem less surprising. If we were to possess an algorithm that could determine whether a Diophantine equation has a solution, we could then simply apply the algorithm to solve problems in the form of those in Theorem 6.7, such as Goldbach's conjecture.

In the remainder of this section we show that the Riemann hypothesis, perhaps the most important and well-known open problem in number theory, is equivalent to the unsolvability of a particular Diophantine equation.[8] As we are not assuming that the reader has previously studied complex analysis, some details will be omitted. We refer the reader to [**Mu**] for an introduction to analytic number theory.

[8] Our equivalence of the Riemann hypothesis is that given by Davis, Matiyasevič, and Robinson in [**DMR**], where they acknowledge the assistance of Harold N. Shapiro (1922–2013).

The *Riemann zeta function* is defined to be

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s},$$

which converges for s with real part greater than 1. Other descriptions of $\zeta(s)$ may be given that converge elsewhere in the complex plane, allowing us to define $\zeta(s)$ for all $s \neq 1$.[9] It turns out that $\zeta(s) = 0$ for $s = -2k$ with $k \in \mathbb{N}$. These are called the *trivial zeroes* of $\zeta(s)$.[10] All other zeroes of $\zeta(s)$ must satisfy $0 < \operatorname{Re} s < 1$.[11] The *Riemann hypothesis*, made by Bernhard Riemann (1826–1866) in 1859, is the following statement:

The real part of each nontrivial zero of $\zeta(s)$ is $\frac{1}{2}$.

The Riemann hypothesis has a strong connection to the distribution of the prime numbers, although the connection is not obvious in its above form. Let $\pi(x)$ be the prime counting function

$$\pi(x) = \sum_{p \leq x} 1.$$

The *prime number theorem*, proved in 1896 independently by Charles Jean de la Vallée Poussin (1866–1962) and Jacques Hadamard (1865–1963), showed that

$$\pi(x) \sim \operatorname{li}(x),$$

where

$$\operatorname{li}(x) = \int_2^x \frac{dt}{\log t}.$$

[9] Using a technique called partial summation, it follows that

$$\zeta(s) = 1 + \frac{1}{s-1} - s \int_1^{\infty} \frac{\{x\}}{x^{s+1}} dx,$$

where $\{x\} = x - \lfloor x \rfloor$ is the fractional part of x. The integral converges for $\operatorname{Re} s > 0$, allowing us to extend $\zeta(s)$ to an analytic function for these values of s, provided $s \neq 1$. Riemann showed how to extend $\zeta(s)$ to the entire complex plane, except for the simple pole at $s = 1$. He also proved the *functional equation*

$$\pi^{-\frac{s}{2}} \Gamma\left(\frac{s}{2}\right) \zeta(s) = \pi^{-\frac{1-s}{2}} \Gamma\left(\frac{1-s}{2}\right) \zeta(1-s).$$

It gives a relation between the value of $\zeta(s)$ at s with the value at $1 - s$.

[10] This can be seen by using the functional equation with the fact that $\Gamma(s)$ has poles at $s = 0, -1, -2, \ldots$ and no zeroes.

[11] The fact that $\zeta(s) \neq 0$ for $\operatorname{Re} s = 1$ is equivalent to the prime number theorem, which is mentioned below. It takes some work to show that this equivalence holds.

Thus the quotient of $\pi(x)$ and $\operatorname{li}(x)$ approaches 1 as x gets large.[12] How closely does $\operatorname{li}(x)$ approximate $\pi(x)$? To answer this question, we look at their difference. It has been shown that

$$|\pi(x) - \operatorname{li}(x)| \leq Axe^{-B\sqrt{\log x}}$$

for certain constants A and B. However, the Riemann hypothesis is equivalent to a stronger bound

$$|\pi(x) - \operatorname{li}(x)| \leq C\sqrt{x}\log x$$

for a certain constant C. In this form, the connection between the Riemann hypothesis and the distribution of the prime numbers is more evident. In fact, it is enough for the above bound to hold for the positive integers in order to imply the Riemann hypothesis. However, it is not clear that the above inequality is a decidable property of x. Given a natural number x, how can we determine if the left-hand side is less than or equal to the right-hand side? We can remove the square root by squaring both sides, but the appearance of $\operatorname{li}(x)$ and $\log x$ present us with some issues from a computational perspective. To avoid these, we will give another form of the Riemann hypothesis, one that replaces $\pi(x)$ with a closely related function.

Theorem 6.8. *For* $\operatorname{Re} s > 1$*, we have*

$$\zeta(s) = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1}.$$

This is called the Euler product for $\zeta(s)$*.*

Proof. Using the geometric series, we have

$$\begin{aligned}\prod_p \left(1 - \frac{1}{p^s}\right)^{-1} &= \prod_p \left(1 + \frac{1}{p^s} + \frac{1}{p^{2s}} + \cdots\right) \\ &= \sum_{n=1}^{\infty} \frac{1}{n^s}.\end{aligned}$$

[12] Using integration by parts or L'Hôpital's rule, one can show that $\operatorname{li}(x) \sim \frac{x}{\log x}$. Thus the prime number theorem is often stated in an equivalent form:

$$\pi(x) \sim \frac{x}{\log x}.$$

The last line is due to the fundamental theorem of arithmetic (Theorem 3.6), which states that every natural number factors uniquely as a product of primes. □

We note that we have not been careful with convergence issues in the above proof and will continue to gloss over such issues in subsequent proofs, as we are not assuming a strong background in analysis on the part of the reader. However, ignoring convergence issues can easily lead one astray, and so much care would have to be taken here in a more rigorous treatment of this material. We refer the reader to [**Mu**] for a formal introduction, which is not essential now for our discussion here.

Theorem 6.9. *For* $\operatorname{Re} s > 1$*, we have*

$$-\frac{\zeta'}{\zeta}(s) = \sum_{n=1}^{\infty} \frac{\Lambda(n)}{n^s},$$

where

$$\Lambda(n) = \begin{cases} \log p & \text{if } n = p^\alpha \text{ for some } \alpha \geq 1, \\ 0 & \text{otherwise,} \end{cases}$$

is the von Mangoldt function.

Proof. Applying the logarithm to the Euler product of Theorem 6.8 and then using the Maclaurin series for $\log(1-x)$, we have

$$\begin{aligned} \log \zeta(s) &= -\sum_p \log\left(1 - \frac{1}{p^s}\right) \\ &= \sum_p \sum_{m=1}^{\infty} \frac{1}{mp^{ms}}. \end{aligned}$$

Differentiating this yields

$$\begin{aligned} \frac{\zeta'}{\zeta}(s) &= -\sum_p \sum_{m=1}^{\infty} \frac{\log p}{p^{ms}} \\ &= -\sum_{n=1}^{\infty} \frac{\Lambda(n)}{n^s}, \end{aligned}$$

as required. □

We now define a function closely related to the prime counting function $\pi(x)$. We let

$$\psi(x) = \sum_{p^\alpha \leq x} \log p = \sum_{n \leq x} \Lambda(n).$$

Thus this function counts all prime powers less than or equal to x, and weights each by $\log p$. The behaviour of $\psi(x)$ and $\pi(x)$ are closely linked. For example, the prime number theorem is equivalent to $\psi(x) \sim x$. It is much easier to work with $\psi(x)$ than $\pi(x)$, partially because it is the sum of the first x numerators in the series for $-(\zeta'/\zeta)(s)$ that was given in Theorem 6.9. We set

$$\psi_1(x) = \sum_{n \leq x} \Lambda(n)(x-n).$$

It can be shown that

$$\psi_1(x) = \int_1^x \psi(t)dt$$

(this is an exercise).[13] The prime number theorem in terms of $\psi_1(x)$ is that $\psi_1(x) \sim \frac{x^2}{2}$.

For $c > 0$, it may be shown that

$$\frac{1}{2\pi i} \int_{(c)} \frac{y^{s+1}}{s(s+1)} ds = \begin{cases} 0 & \text{if } 0 < y < 1, \\ y - 1 & \text{if } y \geq 1, \end{cases} \tag{6.2}$$

although doing so requires contour integration. Here the notation (c) in the integration means we integrate from $c - i\infty$ to $c + i\infty$.[14] Then, taking $c > 1$ so that switching the sum and integral can be justified

[13] Using $\psi_1(x)$ instead of $\psi(x)$ allows us to avoid some convergence issues with a sum over the nontrivial zeroes of $\zeta(s)$.

[14] In particular, we integrate over the line from $c - it$ to $c + it$, and send t to infinity.

and using Theorem 6.9, we have

$$\begin{aligned}\frac{1}{2\pi i}\int_{(c)} -\frac{\zeta'}{\zeta}(s)\frac{x^{s+1}}{s(s+1)}ds &= \frac{1}{2\pi i}\int_{(c)}\sum_{n=1}^{\infty}\frac{\Lambda(n)}{n^s}\frac{x^{s+1}}{s(s+1)}ds\\ &= \sum_{n=1}^{\infty}\frac{n\Lambda(n)}{2\pi i}\int_{(c)}\frac{(x/n)^{s+1}}{s(s+1)}ds\\ &= \sum_{n\le x} n\Lambda(n)\left(\frac{x}{n}-1\right)\\ &= \psi_1(x).\end{aligned}$$

We used (6.2) with $y = x/n$, which allowed us to switch from an infinite sum over all natural numbers to a finite sum over natural numbers up to x. We can also use contour integration to evaluate the above integral by "moving the line of integration to the left". This is a common technique used in complex analysis, but it is outside the scope of this book to cover it in detail. Again, we refer the student to [**Mu**]. We note that $(\zeta'/\zeta)(s)$ has a simple pole at $s = 1$, where $\zeta(s)$ has a pole and has simple poles at the zeroes of $\zeta(s)$. The residue of the pole at $s = 1$ is -1, and the residue of the poles at the trivial zeroes of $\zeta(s)$ is 1. This, with much more justification, yields the following explicit formula for $\psi_1(x)$.

Theorem 6.10. *Let $x \ge 1$. Then*

$$\psi_1(x) = \frac{x^2}{2} - \sum_{\rho}\frac{x^{\rho+1}}{\rho(\rho+1)} - x\log 2\pi + \frac{\zeta'}{\zeta}(-1) + \sum_{k=1}^{\infty}\frac{x^{1-2k}}{2k(1-2k)},$$

where the first sum is over the nontrivial zeroes ρ of $\zeta(s)$ and is taken with multiplicity.[15]

The first term on the right-hand side, which was expected given the prime number theorem, comes from the pole at $s = 1$. The next term comes from the nontrivial zeroes of $\zeta(s)$. The following two terms come from the poles of the integrand at $s = 0$ and $s = -1$.[16] Finally, the last sum comes from the trivial zeroes of $\zeta(s)$.

[15] That is, if ρ were a zero of $\zeta(s)$ of order 2, then it would appear twice in the summation.

[16] We used the fact that $\frac{\zeta'}{\zeta}(0) = \log 2\pi$ here. The reader can find a proof of this in [**Dav**, p. 81].

We are now ready to show an equivalence of the Riemann hypothesis.

Theorem 6.11. *The Riemann hypothesis holds if and only if*

$$\left|\psi_1(x) - \frac{x^2}{2}\right| < 4.564x^{\frac{3}{2}}$$

for every positive natural number x.

Proof. We prove the forward direction first. Suppose the Riemann hypothesis holds, so that every nontrivial zero of $\zeta(s)$ has real part equal to $\frac{1}{2}$. Then Theorem 6.10 yields

$$\left|\psi_1(x) - \frac{x^2}{2}\right| \leq Cx^{\frac{3}{2}}$$

for

$$C = \sum_{\rho} \frac{1}{|\rho||\rho+1|} + \log 2\pi + \left|\frac{\zeta'}{\zeta}(-1)\right| + \sum_{k=1}^{\infty} \frac{1}{2k(2k-1)}.$$

Now, using computational methods, it is possible to show that

$$\left|\frac{\zeta'}{\zeta}(-1)\right| < 1.986.$$

Using results in analytic number theory beyond the scope of this text, one can show[17] that

$$\sum_{\rho} \frac{1}{|\rho||\rho+1|} \leq 2 + \gamma - \log 4\pi.$$

Here, γ is the Euler–Mascheroni constant defined by

$$\gamma = \lim_{n\to\infty} \left(\sum_{k=1}^{n} \frac{1}{k} - \log n\right).$$

[17]Note that since we are assuming the Riemann hypothesis, if ρ is a nontrivial zero of $\zeta(s)$, then $\overline{\rho} = 1 - \rho$. Thus

$$\sum_{\rho} \frac{1}{|\rho||\rho+1|} \leq \sum_{\rho} \frac{1}{|\rho|^2} = \sum_{\rho} \frac{1}{\rho\overline{\rho}} = \sum_{\rho} \frac{1}{\rho(1-\rho)}.$$

To see that the latter sum is equal to $2+\gamma-\log 4\pi$ (which does not require the Riemann hypothesis), let $s = 1$ in the logarithmic derivative of the Hadamard product for

$$s(s-1)\Gamma\left(\frac{s}{2}\right)\zeta(s).$$

We have $0.577 < \gamma < 0.578$. Finally, we have

$$\begin{aligned}\sum_{k=1}^{\infty} \frac{1}{2k(2k-1)} &= \sum_{k=1}^{\infty} \left(\frac{1}{2k-1} - \frac{1}{2k} \right) \\ &= \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \\ &= \log 2.\end{aligned}$$

Thus

$$\begin{aligned}C &\leq 2 + \gamma - \log 4\pi + \log 2\pi + \left| \frac{\zeta'}{\zeta}(-1) \right| + \log 2 \\ &= 2 + \gamma + \left| \frac{\zeta'}{\zeta}(-1) \right| \\ &< 4.564.\end{aligned}$$

We now show the reverse direction. Suppose there is a constant M such that for every positive natural number x, we have

$$\left| \psi_1(x) - \frac{x^2}{2} \right| < M x^{\frac{3}{2}}.$$

Now letting $x \geq 1$ be real, the triangle inequality yields

$$\left| \psi_1(x) - \frac{x^2}{2} \right| \leq |\psi_1(x) - \psi_1(\lfloor x \rfloor)| + \left| \psi_1(\lfloor x \rfloor) - \frac{\lfloor x \rfloor^2}{2} \right| + \left| \frac{\lfloor x \rfloor^2 - x^2}{2} \right|.$$

The second term on the right-hand side is less than $M x^{\frac{3}{2}}$, while for the last term we have

$$\begin{aligned}\left| \frac{\lfloor x \rfloor^2 - x^2}{2} \right| &= \frac{(x - \lfloor x \rfloor)(x + \lfloor x \rfloor)}{2} \\ &< \frac{x + \lfloor x \rfloor}{2} \\ &\leq x \\ &\leq x^{\frac{3}{2}}.\end{aligned}$$

For the first term, we have

$$\begin{aligned}
|\psi_1(x) - \psi_1(\lfloor x \rfloor)| &= \int_{\lfloor x \rfloor}^{x} \psi(t)dt \\
&\leq \psi(x) \\
&= \sum_{n \leq x} \Lambda(n) \\
&\leq \sum_{n \leq x} \log x \\
&= x \log x \\
&< x^{\frac{3}{2}}.
\end{aligned}$$

Thus, for real $x \geq 1$, we have

$$\left| \psi_1(x) - \frac{x^2}{2} \right| < (M+2)x^{\frac{3}{2}}. \tag{6.3}$$

Now, for y a natural number we have

$$\begin{aligned}
\sum_{n \leq y} \frac{\Lambda(n)}{n^s} &= \sum_{n \leq y} \frac{\psi(n) - \psi(n-1)}{n^s} \\
&= \sum_{n \leq y} \frac{\psi(n)}{n^s} - \sum_{n \leq y-1} \frac{\psi(n)}{(n+1)^s} \\
&= \frac{\psi(y)}{y^s} + \sum_{n \leq y-1} \psi(n) \left(\frac{1}{n^s} - \frac{1}{(n+1)^s} \right) \\
&= \frac{\psi(y)}{y^s} + \sum_{n \leq y-1} \psi(n) \int_n^{n+1} \frac{s}{x^{s+1}} dx \\
&= \frac{\psi(y)}{y^s} + s \sum_{n \leq y-1} \int_n^{n+1} \frac{\psi(x)}{x^{s+1}} dx \\
&= \frac{\psi(y)}{y^s} + s \int_1^{y} \frac{\psi(x)}{x^{s+1}} dx \\
&= \frac{\psi(y)}{y^s} + s \frac{\psi_1(y)}{y^{s+1}} + s(s+1) \int_1^{y} \frac{\psi_1(x)}{x^{s+2}} dx.
\end{aligned}$$

To move $\psi(n)$ inside the integral, we used the fact that $\psi(x) = \psi(n)$ for any x with $n \leq x < n+1$. In the last line, we used integration

by parts. Since $\psi(y) \leq y \log y$, sending y to infinity will send the first two terms to 0 when $\operatorname{Re} s > 1$. Using Theorem 6.9 yields

$$-\frac{\zeta'}{\zeta}(s) = s(s+1) \int_1^\infty \frac{\psi_1(x)}{x^{s+2}} dt$$

for $\operatorname{Re} s > 1$. Hence,

$$-\frac{\zeta'}{\zeta}(s) - \frac{s(s+1)}{2(s-1)} = s(s+1) \int_1^\infty \frac{\psi_1(x) - \frac{x^2}{2}}{x^{s+2}} dx.$$

By (6.3) the integral on the right-hand side converges for $\operatorname{Re} s > \frac{1}{2}$. Due to analytic continuation, the equality then holds for all $s \neq 1$ with $\operatorname{Re} s > \frac{1}{2}$, and so $\zeta(s)$ must not vanish for $\frac{1}{2} < \operatorname{Re} s < 1$. This implies the Riemann hypothesis. □

Note that since the reverse direction of the above proof was done using an arbitrary constant M, we may replace the constant 4.564 in Theorem 6.11 with any larger constant, and the equivalence will still hold.

Unfortunately, Theorem 6.11 does not let us use Theorem 6.7 to immediately conclude that the Riemann hypothesis is equivalent to the unsolvability of a Diophantine equation, as it is still not clear that the given property is decidable. We modify the theorem slightly. Let

$$\delta(x) = \prod_{i<x} \prod_{p^\alpha \leq i} p.$$

Theorem 6.12. *The Riemann hypothesis holds if and only if*

$$\left(\sum_{k \leq \delta(x)} \frac{1}{k} - \frac{x^2}{2} \right)^2 < 31x^3$$

for every positive natural number x.

Proof. For x a natural number, we have

$$\log \delta(x) = \sum_{i<x} \sum_{n \leq i} \Lambda(n) = \sum_{n=1}^{x-1} \Lambda(n) \sum_{i=n}^{x-1} 1 = \psi_1(x).$$

A standard estimate from calculus yields

$$\left| \sum_{k \le j} \frac{1}{k} - \log j \right| \le 1.$$

It follows that

$$\left| \sum_{k \le \delta(x)} \frac{1}{k} - \psi_1(x) \right| \le 1. \tag{6.4}$$

Suppose the Riemann hypothesis holds. Then Theorem 6.11 and (6.4) yield

$$\left| \sum_{k \le \delta(x)} \frac{1}{k} - \frac{x^2}{2} \right| \le 5.564 x^{\frac{3}{2}}.$$

Squaring both sides yields the result.

Now suppose

$$\left(\sum_{k \le \delta(x)} \frac{1}{k} - \frac{x^2}{2} \right)^2 < 31 x^3.$$

Taking square roots and using (6.4) yields

$$\left| \psi_1(x) - \frac{x^2}{2} \right| < (\sqrt{31} + 1) x^{\frac{3}{2}}.$$

Since the reverse direction of Theorem 6.11 did not depend on the value of the constant, the theorem with the above bound implies the Riemann hypothesis. □

We are now ready to prove our main theorem on the Riemann hypothesis.

Theorem 6.13. *There is a Diophantine equation that has no solutions if and only if the Riemann hypothesis holds.*

Proof. The form of the Riemann hypothesis given in Theorem 6.12 is clearly a decidable property of the natural numbers x. The result then follows from Theorem 6.7. □

6.4. The Consistency of Axiomatized Theories

In the previous section, we saw that some famous open problems are equivalent to the unsolvability of a Diophantine equation. In this section, we will see that the solvability of Diophantine equations is related to the consistency of axiomatized theories. We will then use Gödel's second incompleteness theorem to draw conclusions about these Diophantine equations.

We can give a listing of the axioms of Peano arithmetic and ZFC set theory. We wrote out the axioms of PA at the beginning of Section 4.4, and the axioms of ZFC were given in Section 2.1. Although it may seem obvious at first that we may algorithmically list the axioms, recall that both theories include axiom schemas, which are actually collections of infinitely many axioms, one for each formula. The induction axiom PA3 of PA and the comprehension and replacement axioms of ZFC are axiom schemas. However, one may give an algorithm to list the formulas of PA and ZFC, and then use this to list the axioms. For example, in ZFC we can begin by listing all axioms but the comprehension and replacement schemas. Then we can use the formula listing algorithm to run through all possible formulas of ZFC, and list the comprehension axiom and the replacement axiom for each formula. By this process, we have an algorithm that will list the axioms of ZFC set theory. A similar process may be carried out for PA. Since there are infinitely many axioms, this algorithm will not terminate. However, each axiom of the theory will eventually appear in our list.

Now we may use our axiom listing algorithm to list the theorems of these theories. Theorems are deduced from previous theorems and axioms by using rules of inference, such as modus ponens (from $P \implies Q$ and P, we may deduce theorem Q). On the odd numbered steps, our theorem listing algorithm uses the axiom listing algorithm to list another axiom. On the even numbered steps, we systematically search through all pairs of previously listed theorems for statements on which we can apply a rule of inference, apply it, and list the resulting theorem. This algorithm creates an infinite list of formulas. If we truncate it at any point, we are left with a proof of the last formula listed. Furthermore, due to the way this list of formulas was

constructed, every formula that has a proof will eventually be listed. Thus this algorithm gives us a list of the theorems of an axiomatized theory.

At first, the existence of a theorem listing algorithm might seem surprising. Why not just run the algorithm and scan the output for interesting new theorems? Unfortunately, most theorems listed will be uninteresting. For example, $\neg(x = x) \implies P$ will be listed for every possible formula P. Furthermore, any interesting theorem listed by the algorithm will include all definitions as part of the theorem, and thus it would be very long. It would take an extreme amount of effort to interpret the theorems listed by the algorithm, and to recognize which are significant. We will see how to use the existence of a theorem listing algorithm to deduce some interesting results. However, the algorithm is not particularly useful for finding new theorems in a practical sense.

We now define a property of the natural numbers. Let $\text{Prop}(n)$ be the statement "there is no contradiction among the first n theorems listed by the theorem listing algorithm". Given a natural number n, we can list the first n theorems and check if both a theorem and its negation appear. Thus $\text{Prop}(n)$ is a decidable property of the natural numbers. By Theorem 6.7, there exists a Diophantine equation that has no solutions if and only if $\text{Prop}(n)$ holds for all natural numbers. On the other hand, $\text{Prop}(n)$ holds for all natural numbers n if and only if the axiomatized theory is consistent. Thus we have the following result.

Theorem 6.14. *There is a Diophantine equation that has no solutions if and only if ZFC set theory is consistent.*

If one were to find solutions to this Diophantine equation, it would follow that ZFC set theory is inconsistent, which would be very bad news and highly unlikely. By Gödel's second incompleteness theorem (Theorem 4.14), ZFC cannot prove its own consistency. Thus, ZFC set theory is not strong enough to prove that this Diophantine equation has no solutions!

We have a similar result for Peano arithmetic:

Theorem 6.15. *There is a Diophantine equation that has no solutions if and only if Peano arithmetic is consistent.*

Again, finding solutions to this Diophantine equation would imply that PA is inconsistent. On the other hand, by Gödel's second incompleteness theorem, PA cannot prove its own consistency. Thus PA is not strong enough to prove that this Diophantine equation has no solution. However, recall that since $\mathbb{N}$ is a model of PA that can be constructed in ZFC, the consistency of PA is a theorem of ZFC. Thus there is a proof in ZFC that this Diophantine equation has no solutions. As discussed at the end of Section 4.4, Gentzen proved that PA is consistent in a theory called "primitive recursive arithmetic augmented with quantifier-free transfinite induction up to the ordinal ϵ_0". Thus there is a proof in this theory that the Diophantine equation of Theorem 6.15 has no solution. To prove this equation has no solutions, a proof technique unavailable in PA must be used, such as transfinite induction on infinite ordinals. It is surprising that such strong techniques must be used to show that some Diophantine equations have no solutions.

Let us return to ZFC set theory. A cardinal κ is called *inaccessible*[18] if it satisfies the following three properties:

- κ is not the limit of a sequence of less than κ cardinals, each less than κ.[19]
- For all $\lambda < \kappa$, $2^\lambda < \kappa$.
- κ is uncountable.

Cardinals that satisfy the first property are called *regular* cardinals. Cardinals that are not regular are called *singular*. For example, consider the sequence

$$\aleph_0, \aleph_1, \aleph_2, \ldots,$$

where the subscript ranges over all elements of the ordinal ω. The limit of this sequence is the cardinal $\aleph_\omega$. Since $|\omega| = \aleph_0$, $\aleph_\omega$ is the

[18] Often these cardinals are called *strongly inaccessible*, to distinguish them from the related *weakly inaccessible* cardinals. A weakly inaccessible cardinal is a limit cardinal that satisfies the first and third inaccessible cardinal properties.

[19] Alternatively, κ is not the (cardinal) sum of less than κ cardinals, each less than κ. Or, κ is not the union of less than κ cardinals, each less than κ.

limit of a sequence of less than $\aleph_\omega$ cardinals, each is less than $\aleph_\omega$. Thus $\aleph_\omega$ is a singular cardinal.

The first property of an inaccessible cardinal shows that, in a sense, the cardinal cannot be broken into smaller pieces. The second property shows that it cannot be reached from below by exponentiation.[20] Note that since every cardinal less than $\aleph_0$ is finite, $\aleph_0$ satisfies these first two properties. Thus an inaccessible cardinal is an uncountable cardinal that, when compared to preceding cardinals, behaves in a way similar to the way $\aleph_0$ behaves when compared to the finite cardinals. The first infinite cardinal $\aleph_0$ is a countable cardinal that is "far above" the preceding finite cardinals. Similarly, an inaccessible cardinal is an uncountable cardinal that is "far above" the preceding cardinals.

Suppose we were able to prove the existence of an inaccessible cardinal κ using ZFC set theory. Using material beyond the scope of this text, one can then use κ to construct a model of ZFC set theory, which implies that ZFC is consistent. However, this contradicts Gödel's second incompleteness theorem, which asserts that ZFC cannot prove its own consistency. Thus the existence of an inaccessible cardinal cannot be proven from the axioms of ZFC. We may add to ZFC the axiom "there exists an inaccessible cardinal". However, this results in a stronger system, which we will call ZFC^+. ZFC^+ can prove all theorems of ZFC, but it can also prove that ZFC is consistent, which ZFC itself cannot prove. Of course, one may ask if ZFC^+ is consistent. By Gödel's second incompleteness theorem, ZFC^+ cannot prove its own consistency.

Consider the Diophantine equation of Theorem 6.14. This Diophantine equation has no solutions if and only if ZFC is consistent. We previously concluded that ZFC cannot prove that this Diophantine equation has no solutions. Since ZFC^+ can prove that ZFC is consistent, we can prove that this Diophantine equation has no solutions in ZFC^+. Here is a Diophantine equation that makes use of a statement as strong as "there exists an inaccessible cardinal" in order to prove that it has no solutions!

[20] Equivalently, the second property shows that the cardinal, as a set, is closed under exponentiation.

Since we may construct an algorithm to list the theorems of ZFC^+, there is a Diophantine equation that has no solutions if and only if ZFC^+ is consistent. This Diophantine equation would need something stronger than what is available in ZFC^+ in order to show that it has no solutions!

Exercises

6.1. Suppose every even number greater than 2 can be written as the sum of two primes. Prove that every odd number greater than 5 can be written as the sum of three primes.

6.2. If natural numbers p and $p+2$ are both prime, they are called *twin primes*. It is conjectured that there are infinitely many pairs of twin primes. It is an open question if the twin prime conjecture is equivalent to a decidable property holding for all natural numbers. Let us make a *strong twin prime conjecture*: For every natural number n, there exist twin primes p and $p+2$ with $n \leq p \leq 2^{2^n}$". Show that the strong twin prime conjecture implies the twin prime conjecture, and that it is equivalent to the unsolvability of a Diophantine equation.

6.3. Show that

$$\sum_{d|n} \Lambda(d) = \log n.$$

6.4. Recall that

$$\psi(x) = \sum_{p^\alpha \leq x} \log p = \sum_{n \leq x} \Lambda(n).$$

Prove that the lowest common multiple of the numbers 1, 2, ..., n is equal to $e^{\psi(n)}$.

6.5. Observe that every prime between n and $2n$ divides $\binom{2n}{n}$. Since the latter is a coefficient in the binomial expansion of $(1+1)^{2n}$, deduce that

$$\sum_{n<p\leq 2n} \log p \leq 2n \log 2.$$

Let $n = 2^r$ and sum for $r = 0, 1, \ldots, m$ to deduce

$$\sum_{p \le 2^{m+1}} \log p \le 2^{m+2} \log 2.$$

Finally, given n, take m with $2^m < n \le 2^{m+1}$ to deduce

$$\sum_{p \le n} \log p \le 4n \log 2.$$

6.6. Observe that

$$\sum_{\substack{p^\alpha \le x \\ \alpha \ge 2}} \log p = \sum_{p \le \sqrt{x}} \log p \sum_{2 \le \alpha \le \frac{\log x}{\log p}} 1.$$

Use the crude bound $\pi(x) \le x$ to show that the above sum is at most $\sqrt{x} \log x$, which is bounded above by x. Use this and the previous exercise to deduce that $\psi(x) \le (4 \log 2 + 1)x$.

6.7. Let $x \ge 1$ be real. Show that

$$\int_1^x \psi(t) dt = \sum_{n \le x} \Lambda(n)(x - n).$$

6.8. Show that

$$\left| \sum_{k \le j} \frac{1}{k} - \log j \right| \le 1.$$

6.9. A cardinal κ is singular if it is the limit of a sequence of less than κ cardinals, each less than κ. Show that this is equivalent to each of the following statements.

(a) κ is the sum of less than κ cardinals, each less than κ.

(b) κ is the union of less than κ cardinals, each less than κ.

6.10. In the Exercises for Chapter 1, it was shown that the union of countably many countable sets is countable. Recall that the cardinal successor of $\aleph_0$, called $\aleph_1$, is the first uncountable cardinal. Show that $\aleph_1$ is a regular cardinal.

Chapter 7

Hilbert's Tenth Problem over Number Fields

7.1. Background on Algebraic Number Theory

Recall that a complex number α is called an *algebraic number* if there is a polynomial

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

with each a_i rational and $a_n \neq 0$ such that $f(\alpha) = 0$. In other words, an algebraic number is the root of a polynomial with rational coefficients. An *algebraic integer* is the root of a *monic* polynomial with integer coefficients (that is, $a_n = 1$). Thus, all algebraic integers are algebraic numbers but not conversely. For example, the reader may easily verify that $\sqrt{2}/3$ is an algebraic number but not an algebraic integer. Recall that a complex number that is not an algebraic number is called *transcendental*. We saw in Chapter 1 that algebraic numbers comprise a countable set so that the set of transcendental numbers is uncountable. This was proved by Cantor in 1874.

Though the set of transcendental numbers is an uncountable set, it is often difficult to exhibit an explicit example. Liouville produced

the first example in 1853. He showed that

$$\sum_{n=1}^{\infty} \frac{1}{2^{n!}}$$

is a transcendental number. We refer the student to the recent book [**MR**] for a gentle introduction to this fascinating topic. A good resource for self-instruction in algebraic number theory is the problems book [**ME**].

If we let $\mathbb{Q}[x]$ denote the polynomial ring over the rational numbers and let α be an algebraic number, then the set

$$S = \{f(\alpha) : f \in \mathbb{Q}[x]\}$$

can be made into a field under the usual addition and multiplication. Indeed, as α satisfies a polynomial equation with rational coefficients, we can speak of its *minimal polynomial* defined as the monic polynomial of least degree (necessarily unique by the Euclidean algorithm). Then if $p(x)$ is the minimal polynomial of α, it is easy to see that

$$S \simeq \mathbb{Q}[x]/(p(x)).$$

Moreover, as $p(x)$ is the polynomial of minimal degree, it is necessarily irreducible. Consequently, the ideal $(p(x))$ is a prime ideal, and the quotient

$$\mathbb{Q}[x]/(p(x))$$

is a field. We denote this field as $\mathbb{Q}(\alpha)$. Viewed as a vector space over $\mathbb{Q}$, we see that it has finite dimension. If n is the degree of the minimal polynomial of α, then

$$1, \alpha, \alpha^2, \ldots, \alpha^{n-1}$$

is a basis of $\mathbb{Q}(\alpha)$ over $\mathbb{Q}$.

A subfield $K \subseteq \mathbb{C}$ is called an *algebraic number field* if its dimension over $\mathbb{Q}$ is finite. This dimension is called the *degree* of K and is denoted $[K : \mathbb{Q}]$. Thus, if the minimal polynomial of α has degree n, then $\mathbb{Q}(\alpha)$ has degree n over $\mathbb{Q}$. All algebraic number fields are of the form $\mathbb{Q}(\alpha)$ for some α though this is not a trivial fact. It follows from the *theorem of the primitive element*. This theorem says that if α and β are algebraic numbers and $\mathbb{Q}(\alpha, \beta)$ is the field generated by α and β, then there is an algebraic number θ such that $\mathbb{Q}(\alpha, \beta) = \mathbb{Q}(\theta)$.

If α is an algebraic number and $p(\alpha)$ is its minimal polynomial of degree n, we can list the roots of $p(x)$ as

$$\alpha = \alpha^{(1)}, \alpha^{(2)}, \ldots \alpha^{(n)}.$$

The fields $\mathbb{Q}(\alpha^{(i)})$, $1 \leq i \leq n$ are called *conjugate fields*. If $K = \mathbb{Q}(\alpha)$, the map $\alpha \mapsto \alpha^{(i)}$ extends to a monomorphism of fields $\mathbb{Q}(\alpha) \to \mathbb{Q}(\alpha^{(i)})$, and we refer to these maps as *embeddings* of K into $\mathbb{C}$. If *all* conjugate fields of K are identical to K, then K is called a *normal* (or *Galois*) extension of $\mathbb{Q}$. For example, $\mathbb{Q}(\sqrt[3]{2})$ is not a Galois extension because the conjugates of $\sqrt[3]{2}$ are $\sqrt[3]{2}$, $\omega\sqrt[3]{2}$, and $\omega^2\sqrt[3]{2}$, where $\omega = e^{2\pi i/3}$ is a primitive cube root of unity, which is not a real number. Indeed, if $\mathbb{Q}(\sqrt[3]{2}) = \mathbb{Q}(\omega\sqrt[3]{2})$, then $\omega \in \mathbb{Q}(\sqrt[3]{2})$, which is a contradiction.

Given a number field K, we consider the set

$$\mathcal{O}_K := \{\alpha \in K : \alpha \text{ is an algebraic integer}\}$$

consisting of all algebraic integers of K. Using basic algebra, one shows that $\mathcal{O}_K$ is a ring under the usual operations of addition and multiplication, a fact first established by Richard Dedekind (1831–1916). We call $\mathcal{O}_K$ the ring of integers of K. The invertible elements of $\mathcal{O}_K$, denoted $\mathcal{O}_K^\times$, are called *units*.

It seems appropriate to include here a brief sketch of the life of Richard Dedekind since he was a contemporary of Cantor who also realized the importance of infinity as a mathematical idea. The *constructionists* spearheaded by Leopold Kronecker (1823–1891) were opposed to Cantor's revolutionary ideas regarding infinities. But Dedekind was not. In fact, Cantor found a sympathetic friend in Dedekind and there was considerable correspondence between them regarding this new development of mathematics. Many of the foundational ideas of algebraic number theory we are discussing here are due to Dedekind and one of the important questions Dedekind had related to the axiomatic construction of the real numbers.

Being the last doctoral student of the famous C. F. Gauss, Dedekind was a colleague of both Riemann and Dirichlet. One question that occupied Dedekind's attention was how to construct the real numbers. As we saw in an earlier chapter, the natural numbers, the

integers, and even the rational numbers can be constructed inductively using set theory. Dedekind wondered if a similar idea could be used to construct the real numbers. This led him to the notion of a *Dedekind cut*, which we now describe. Dedekind first noticed that any rational number r divides the set of rational numbers into two disjoint sets, namely those that are less than r and those that are greater than r. He then realized that we can define a real number using the set of rational numbers combined with this order relation. More precisely, a real number x is a *cut* of the rational numbers into two disjoint sets A and B such that every element of A is less than every element of B. One can then define the usual operations of addition and multiplication of real numbers using sets. For example, the irrational number $\sqrt{2}$ can be described as the cut (A, B) with A consisting of all rational numbers r with $r^2 < 2$ and B consisting of rational numbers whose square is greater than 2. In his inspired *précis* of the lives of famous mathematicians, Hawking wrote that in 1874, Dedekind met Cantor for the first time and that "it is through Cantor's correspondence with Dedekind in the 1870s and 1880s that historians have come to understand the development of Cantor's ideas."[1]

It is worth mentioning here that an alternative construction of the real numbers based on the notion of distance was given by Augustin-Louis Cauchy (1789–1857). He viewed real numbers as limits of convergent sequences of rational numbers. This viewpoint has the advantage of considerable generalization where an order relation may not exist, but a metric does. For instance, this facilitates the construction of the p-adic numbers, and one then sees the real numbers as a *spectral line* in a wide spectrum of fields constructed from the rational numbers. We will not go into too much detail here but refer the interested reader to [**Mu**, Chapter 10].

Dedekind's interest in understanding infinite sets and how to work with them was partially motivated by his discovery of the theory of ideals in an effort to understand factorization of numbers. Though there are structural similarities between $\mathbb{Z}$ and $\mathcal{O}_K$ in that they are

[1] See [**Haw**, p. 1070]. Unfortunately, this précis of Hawking has serious typographical errors. The description of the Dedekind cut for $\sqrt{2}$ on page 1069 is erroneous. In both sets A_1 and A_2, one must replace $\sqrt{2}$ by 2 which is the key point!

both examples of rings, he realized that the unique factorization property of the ring $\mathbb{Z}$ does not extend in general to $\mathcal{O}_K$. This was realized by many mathematicians in their early attempts to solve Fermat's last theorem.[2] For example, the ring of integers of the field $\mathbb{Q}(\sqrt{-5})$ is $\mathbb{Z}[\sqrt{-5}]$. In this ring, one has two factorizations of 6 as

$$6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5}).$$

One can show that none of $2, 3, 1 \pm \sqrt{-5}$ are units and they are all irreducible elements in the sense that they cannot be factored further.

The discovery of nonunique factorization for many of the rings $\mathcal{O}_K$ led to a search for a general theory of divisibility. It led to fundamental questions. This intellectual ferment gave rise to two theories of divisibility, one formulated by Kronecker and another by Dedekind. Both are essentially the same, but Dedekind's theory deftly used the notion of infinity, a notion and method challenged by Kronecker. This fascinating chapter in the evolution of algebraic number theory is illustrated by Stillwell in the reprint of Dedekind's classic *Theory of Algebraic Numbers* [**De**].

Dedekind essentially created the theory of modules to understand divisibility. For instance, the familiar idea that a natural number d divides a number n is equivalently formulated using infinite sets,

$$d \mid n \;\Leftrightarrow\; n\mathbb{Z} \subseteq d\mathbb{Z},$$

which led to the aphorism "to contain is to divide". When we reflect upon Dedekind's other contributions to mathematics, specifically to our understanding of a real number via Dedekind cuts, we find this notion of divisibility as natural. From this, Dedekind was led to formulate his general theory of ideals. He discovered that for the rings $\mathcal{O}_K$, one can recover from the failure of unique factorization if one moves into the realm of ideals. Prime numbers are then replaced by prime ideals, and Dedekind proved that in $\mathcal{O}_K$ every ideal can be factored uniquely as a product of prime ideals. Since $\mathbb{Z}$ is a principal ideal domain, we then see that Dedekind's theorem is the natural generalization of the classical unique factorization theorem.

[2]Fermat asserted that for $n \geq 3$, $x^n + y^n = z^n$ has no nontrivial solutions in the integers, a fact that took 358 years to prove.

Dedekind also discovered a natural generalization of the Riemann zeta function. He proved that every nonzero ideal $\mathfrak{a}$ of $\mathcal{O}_K$ has finite index, and he defined the *norm* of $\mathfrak{a}$, denoted $\mathrm{N}(\mathfrak{a})$, as the index $[\mathcal{O}_K : \mathfrak{a}]$. Using his theory of ideals, he showed this was a multiplicative function of ideals and introduced what we now call the *Dedekind zeta function*,

$$\zeta_K(s) = \sum_{\mathfrak{a} \neq 0} \mathrm{N}(\mathfrak{a})^{-s},$$

where the sum is over all nonzero ideals of $\mathcal{O}_K$. By using his unique factorization theorem, he showed that $\zeta_K(s)$ admits an *Euler product*

$$\zeta_K(s) = \prod_{\mathfrak{p}} \left(1 - \frac{1}{\mathrm{N}(\mathfrak{p})^s}\right)^{-1},$$

where the product runs over nonzero prime ideals of $\mathcal{O}_K$. One can show that both the series and product converge absolutely for $\mathrm{Re}(s) > 1$. In analogy with the Riemann zeta function, Dedekind conjectured that $\zeta_K(s)$ extends to an analytic function for all $s \in \mathbb{C}$ except $s = 1$ where it has a simple pole. This was later proved by Erich Hecke (1887–1947) in 1918.[3]

If the field K is $\mathbb{Q}(\alpha)$, then we already spoke of its conjugate fields. Since α satisfies a polynomial equation with rational coefficients, we see that if β is a conjugate of α, so is $\overline{\beta}$, which is the complex conjugate. Thus we may pair up the nonreal conjugates of α in conjugate pairs. In other words, if α has degree n, we may list its conjugates as

$$\alpha_1, \alpha_2, \ldots, \alpha_{r_1}, \alpha_{r_1+1}, \ldots, \alpha_{r_1+r_2}, \overline{\alpha}_{r_1+1}, \ldots, \overline{\alpha}_{r_1+r_2},$$

where the number of *real* conjugates of α is r_1 and the number of nonreal (or *complex*) conjugates is $2r_2$. Clearly, $r_1 + 2r_2 = n$. We speak of the maps

$$K \ \rightarrow \ \mathbb{Q}(\alpha^{(i)})$$

with $\alpha^{(i)} = \alpha_i$, $1 \leq i \leq r_1$ as the real embeddings of K and the maps

$$K \ \rightarrow \ \mathbb{Q}(\alpha_i)$$

with $r_1 + 1 \leq i \leq r_1 + r_2$ as the nonreal (or complex) embeddings.

[3]The *generalized Riemann hypothesis*, an open problem, is that all nontrivial zeros of $\zeta_K(s)$ have a real part of $1/2$. In [**Fo**] the methods of Theorem 6.13 were extended to show that, for any given computably enumerable collection of number fields, the generalized Riemann hypothesis holds for each of these number fields if and only if a particular Diophantine equation has no solution.

The structure of the *unit group* $\mathcal{O}_K^\times$ was determined by Gustav Lejeune Dirichlet (1805–1859) in 1846 although he did not formulate his results in precisely these terms. He proved that

$$\mathcal{O}_K^\times \simeq W \times \mathbb{Z}^r$$

with $r = r_1 + r_2 - 1$ and W being a finite group consisting of roots of unity. In other words, the unit group of $\mathcal{O}_K$ is a finitely generated group of rank r, with its torsion subgroups being a finite group of roots of unity.

We say a field K is *totally real* if all its conjugate fields are real. Thus, in such a case, the unit rank is $n-1$ where $n = [K : \mathbb{Q}]$, because $r_2 = 0$. In other words, the unit rank is as large as possible only in the case where K is totally real.

The structure of $\mathcal{O}_K$ was first studied in detail by Dedekind, who single handedly developed the theory of modules over $\mathbb{Z}$ for this purpose. Stillwell [**De**] writes

> Dedekind's invention of ideals in the 1870s was a major turning point in the development of algebra. His aim was to apply ideals to number theory, but to do this he had to build the whole framework of commutative algebra: fields, rings, modules, and vector spaces. These concepts, together with groups, were to form the core of the future abstract algebra. At the same time, he created algebraic number theory, which became the temporary home of algebra while its core concepts were growing up.[4]

These are concepts we now take for granted but 150 years ago, they were nonexistent.

Dedekind thus began his study of rings of integers of number fields. He showed that $\mathcal{O}_K$ is a finitely generated $\mathbb{Z}$-module of rank n. In other words, there exist algebraic integers $\omega_1, \omega_2, \ldots, \omega_n$ in $\mathcal{O}_K$ such that

$$\mathcal{O}_K = \mathbb{Z}\omega_1 \oplus \cdots \oplus \mathbb{Z}\omega_n.$$

[4] See [**De**, p. 3].

That is, every algebraic integer of K is uniquely expressible as a $\mathbb{Z}$-linear combination of $\omega_1, \omega_2, \ldots, \omega_n$. He called the set of these numbers an *integral basis* for $\mathcal{O}_K$. Such a basis is not unique and there are many choices for it. However, Dedekind proved that if Ω is the $n \times n$ matrix whose (i,j)-th entry is $\omega_i^{(j)}$, then the quantity

$$(\det(\Omega))^2$$

does not depend on the choice of basis and is in fact in $\mathbb{Z}$. He denoted this integer as d_K and called it the *discriminant* of K.

Since every ideal of $\mathcal{O}_K$ has a unique factorization into prime ideals, we may consider for each prime p the principal ideal $p\mathcal{O}_K$ of $\mathcal{O}_K$. This admits a unique factorization,

$$p\mathcal{O}_K = \mathfrak{p}_1^{e_1} \cdots \mathfrak{p}_g^{e_g},$$

where the $\mathfrak{p}_i$ are distinct prime ideals of $\mathcal{O}_K$. If some $e_i > 1$, we say that p *ramifies* in K. Dedekind proved that p ramifies in K if and only if $p \mid d_K$. Thus, there are only finitely many ramified primes and, in some sense, d_K measures the extent of ramification.

The research of Dedekind inspired Hermann Minkowski (1864–1909) in the 1890s to develop the theory of the geometry of numbers. Using his theory, Minkowski showed that $|d_K| > 1$ if $K \neq \mathbb{Q}$. In other words, for any nontrivial extension of $\mathbb{Q}$, there exist ramified primes.

Dedekind's theory led to further reflections on the notion of a prime. Indeed, if $\mathfrak{p}$ is a prime ideal of $\mathcal{O}_K$, we can define a *valuation*, denoted $v_\mathfrak{p}$ as follows. For any $\alpha \in \mathcal{O}_K$, the principal ideal (α) admits a unique factorization into prime ideals,

$$(\alpha) = \prod_{\mathfrak{p}} \mathfrak{p}^{v_\mathfrak{p}(\alpha)},$$

where the product is over *all* prime ideals of $\mathcal{O}_K$ and for any fixed α, $v_\mathfrak{p}(\alpha) = 0$ except for finitely many $\mathfrak{p}$. In this way, we obtain a function

$$v_\mathfrak{p} : \mathcal{O}_K \to \mathbb{R}_+ \cup \{0\},$$

where $\mathbb{R}_+$ denotes the positive reals. Clearly, we may extend this definition to K since K is the field of fractions of $\mathcal{O}_K$. That is, if $\alpha/\beta \in K$, $\beta \neq 0$ with $\alpha, \beta \in \mathcal{O}_K$, we may set $v_\mathfrak{p}(\alpha/\beta) = v_\mathfrak{p}(\alpha) - v_\mathfrak{p}(\beta)$.

Let us now define[5]

$$|x|_{\mathfrak{p}} := \mathrm{N}(\mathfrak{p})^{-v_{\mathfrak{p}}(\alpha)}.$$

Thus, $|\cdot|_{\mathfrak{p}}$ is an example of a *valuation* (synonymously called *place* or *norm* or *absolute value*) which is defined as a map

$$v : K \to \mathbb{R}_+ \cup \{0\}$$

that satisfies the following three properties:

(1) $v(0) = 0$, and $v(x) = 0$ if and only if $x = 0$;

(2) $v(xy) = v(x)v(y)$ for all $x, y \in K$;

(3) there is a constant C such that

$$v(x + y) \le C \max(v(x), v(y))$$

for all $x, y \in K$.

The reader can verify that $v_{\mathfrak{p}}$ satisfies these properties with $C = 1$. The familiar absolute value also satisfies these properties with $C = 2$ as a consequence of the triangle inequality. Moreover, we can define for each embedding $K \to K^{(i)}$, a valuation $v^{(i)}$,

$$\alpha \mapsto |\alpha^{(i)}|, \qquad 1 \le i \le n.$$

These valuations are called Archimedean valuations. The others, namely $|\cdot|_{\mathfrak{p}}$, are called non-Archimedean valuations.

There is always the trivial valuation, $v(0) = 0$ and $v(x) = 1$ for all $x \neq 0$. We usually discard this trivial valuation in our study. Each valuation defines a metric on K that makes K into a Hausdorff topological field in which the field operations are continuous. Two valuations are said to be *equivalent* if they induce the same topology on K. A celebrated theorem of Alexander Ostrowski (1893–1986), proved in 1918, states that up to equivalence, any nontrivial valuation of K is either $|\cdot|_{\mathfrak{p}}$ corresponding to a prime ideal $\mathfrak{p}$ of $\mathcal{O}_K$ or $v^{(i)}$ for some embedding of K. It is this theorem that led to the view that valuations are to be seen as the appropriate generalization of the concept of a prime. This perspective gave rise to our modern *adelic viewpoint* that has transformed number theory and mathematics in general. The

[5]The use of $\mathrm{N}(\mathfrak{p})$ here can be replaced by any constant $c > 1$, and this would give rise to the same topology on K. We have used here a convenient normalization often used in algebraic number theory so as to make formulas (such as the product formula) more elegant.

modern perspective is to view an algebraic number field K as a spectrum of valuations consisting of both non-Archimedean (finite) and Archimedean (infinite) places. This has served as a grand unifying theme in number theory culminating in Tate's thesis. Briefly stated, Tate's thesis takes the adelic perspective to study the Dedekind zeta function and to derive its analytic continuation and functional equation as a consequence of Fourier duality.

The search for a generalization of Dirichlet's theorem on the infinitude of primes in arithmetic progressions led to the study of *ideal class groups* and *generalized ideal class groups*. The initial impulse is to study the *ideal class group* defined as the multiplicative group generated by the ideals of $\mathcal{O}_K$ modulo the subgroup of principal ideals. In 1871 Dedekind proved that this group is a finite group. The ideal class group is a measurement of how far $\mathcal{O}_K$ is from being a principal ideal domain.

The generalized ideal class groups are defined as follows. Let $\mathfrak{m}_0$ be an ideal of $\mathcal{O}_K$, and let $\mathfrak{m}_\infty$ be a subset of real embeddings of K. Let I be the multiplicative group generated by all ideals of $\mathcal{O}_K$ coprime to $\mathfrak{m}_0$. We consider the subgroup P of fractional ideals (α) (that is, the multiplicative group generated by principal ideals) with $\alpha \equiv 1 \pmod{\mathfrak{m}_0}$ *and* $\alpha^{(i)} > 0$ *for all* $i \in \mathfrak{m}_\infty$. The *generalized ideal class group* determined by the pair $(\mathfrak{m}_0, \mathfrak{m}_\infty)$ is the quotient I/P. These groups are viewed as the number field generalizations of the more familiar coprime residue classes of the rational number field. Indeed, if $m_0 \in \mathbb{Z}$ and m_∞ is the usual embedding $\mathbb{Q} \hookrightarrow \mathbb{R}$, then the ideal class group determined by (m_0, m_∞) is the usual group $(\mathbb{Z}/m_0\mathbb{Z})^\times$. Thus, the ideal classes of the generalized ideal class groups are the natural generalizations of arithmetic progressions.

Once these ideal class groups are defined, it is then natural to extend Dedekind's theorem and show that they are finite. Then, to extend Dirichlet's theorem on the infinitude of primes in arithmetic progressions, we define the L-series $L(s, \chi)$ for each character χ of the generalized ideal class group

$$L(s, \chi) = \sum_{\mathfrak{a}} \frac{\chi(\mathfrak{a})}{(\mathrm{N}\,\mathfrak{a})^s}.$$

These are generalizations of the Dedekind zeta function. In 1918, Hecke proved that these L-series extend to the entire complex plane and satisfy a suitable functional equation. Consequently, we refer to these L-series as Hecke L-series.

In the years following the work of Hecke, algebraic number theory underwent a remarkable transformation, largely through the use of the topology of adele rings of number fields. These rings are defined as follows.

As noted earlier, a convenient way of looking at prime ideals of $\mathcal{O}_K$ is to view them as part of the spectrum of valuations of the field K. If v is a *valuation*, then we may complete K to get the v-adic field K_v. If v is non-Archimedean (that is, corresponding to a prime ideal of $\mathcal{O}_K$), then we define the ring of *v-adic integers* to be the set of $x \in K_v$ such that $v(x) \leq 1$. We denote this ring as $\mathcal{O}_v$. The adele ring $\mathbb{A}_K$ consists of infinite tuples (x_v), as v ranges over all the inequivalent valuations (or places) with $x_v \in K_v$, but with the proviso that $x_v \in \mathcal{O}_v$ for all but finitely many places. Addition and multiplication in this ring are defined componentwise. This makes $\mathbb{A}_K$ into a commutative ring. We make it a *locally compact* topological ring by declaring that for each finite set S of places containing the Archimedean places, the set

$$\prod_{v \in S} K_v \times \prod_{v \notin S} \mathcal{O}_v,$$

with the product topology, is a basic neighbourhood of the identity element of $\mathbb{A}_K$. One can view K as embedded into $\mathbb{A}_K$ via the map

$$x \mapsto (x, x, x, \ldots).$$

If $\mathrm{GL}_n(\mathbb{A}_K)$ is the group of invertible $n \times n$ matrices with entries in $\mathbb{A}_K$, one can similarly define a topology on it so as to make it into a locally compact topological group. Indeed, for each finite set S containing the infinite places, we declare

$$\prod_{v \in S} \mathrm{GL}_n(K_v) \times \prod_{v \notin S} \mathrm{GL}_n(\mathcal{O}_v),$$

with the product topology, as a basic neighbourhood of the identity.

The group of units of $\mathbb{A}_K$, denoted $\mathbb{A}_K^\times$, is called the *group of ideles*. It is clear that $K^\times$ is embedded in $\mathbb{A}_K^\times$, as before. The group

$\mathbb{A}_K^\times/K^\times$ is called the *idele class group*. Characters of this group are called *grossencharacters*, and Hecke's work shows that one can associate L-functions to these grossencharacters which extend to the entire complex plane and satisfy a suitable functional equation. This formulation of Hecke's work is the essence of Tate's thesis.

7.2. Introduction to Zeta Functions and L-functions

All of the characters attached to generalized ideal class groups turn out to be special cases of grossencharacters. Robert Langlands (born 1936) formulated a vast generalization of the construction of these L-functions.

As hinted before, $\mathrm{GL}_n(\mathbb{A}_K)$ is a locally compact topological group in which $\mathrm{GL}_n(K)$ is embedded diagonally as a discrete subgroup. If I denotes the $n \times n$ identity matrix and

$$Z := \{zI : z \in \mathbb{A}_K^\times\},$$

then the coset space $Z\,\mathrm{GL}_n(K)\backslash \mathrm{GL}_n(\mathbb{A}_K)$ has finite volume with respect to any $\mathrm{GL}_n(\mathbb{A}_K)$ invariant measure.

Let us now fix a grossencharacter ω and consider the Hilbert space

$$L^2(\mathrm{GL}_n(K)\backslash \mathrm{GL}_n(\mathbb{A}_K), \omega)$$

consisting of measurable functions ϕ on $\mathrm{GL}_n(K)\backslash \mathrm{GL}_n(\mathbb{A}_K)$ satisfying

(i) $\phi(zg) = \omega(z)\phi(g) \qquad \forall z \in Z, g \in \mathrm{GL}_n(K)\backslash \mathrm{GL}_n(\mathbb{A}_K)$;

(ii) $\displaystyle\int_{Z\,\mathrm{GL}_n(K)\backslash \mathrm{GL}_n(\mathbb{A}_K)} |\phi(g)|^2 dg < \infty.$

A special role is played by the *standard parabolic subgroups* of GL_n. They are in one-to-one correspondence with partitions

$$n = n_1 + n_2 + \cdots + n_r.$$

If M_n is a generic $n \times n$ matrix with entries in $\mathbb{A}_K$ and I_n is the $n \times n$ identity matrix, the *standard parabolic subgroup* consists of matrices

of the form

$$\begin{pmatrix} M_{n_1} & * & \cdots & * \\ & M_{n_2} & \cdots & * \\ & & \ddots & \\ & & & M_{n_r} \end{pmatrix},$$

and any subgroup conjugate to a standard parabolic subgroup is called a *parabolic subgroup*. Such a subgroup P admits a decomposition (called the *Levi decomposition*) of the form

$$P = MN,$$

where N is the *unipotent radical* (consisting of elements of P, all of whose eigenvalues are equal to 1).

The *subspace of cusp forms*

$$L^2_o(\mathrm{GL}_n(K)\backslash \mathrm{GL}_n(\mathbb{A}_K), \omega)$$

of $L^2(\mathrm{GL}_n(K)\backslash \mathrm{GL}_n(\mathbb{A}_K), \omega)$ is defined as the collection of ϕ satisfying (i) and (ii) above *and* for all parabolic subgroups P of $\mathrm{GL}_n(\mathbb{A}_K)$, we have

$$\text{(iii)} \int_{\mathrm{N}(K)\backslash \mathrm{N}(\mathbb{A}_K)} \phi(ng)dn = 0 \qquad \forall g \in \mathrm{GL}_n(\mathbb{A}_K),$$

where N is the unipotent radical of P.

The *right regular representation* R of $\mathrm{GL}_n(\mathbb{A}_K)$ is defined as an action of $\mathrm{GL}_n(\mathbb{A}_K)$ on

$$L^2(\mathrm{GL}_n(K)\backslash \mathrm{GL}_n(\mathbb{A}_K), \omega)$$

by the formula

$$(R(g)\phi)(x) = \phi(xg) \qquad \forall \phi \in L^2(\mathrm{GL}_n(K)\backslash \mathrm{GL}_n(\mathbb{A}_K), \omega)$$

and $x, g \in \mathrm{GL}_n(\mathbb{A}_K)$. An *automorphic representation* is a subquotient of the right regular representation. A *cuspidal automorphic representation* is a subrepresentation of the right representation of $\mathrm{GL}_n(\mathbb{A}_K)$ on

$$L^2_o(\mathrm{GL}_n(K)\backslash \mathrm{GL}_n(\mathbb{A}_K), \omega).$$

It is called *admissible* if its restriction to the compact group

$$\prod_{v \text{ complex}} U_n(\mathbb{C}) \times \prod_{v \text{ real}} \mathcal{O}_n(\mathbb{R}) \times \prod_{v \text{ finite}} \mathrm{GL}_n(\mathcal{O}_v)$$

contains each irreducible representation with finite multiplicity (where the product is over all valuations of K, and where $U_n(\mathbb{C})$, $\mathcal{O}_n(\mathbb{R})$ denote the unitary and orthogonal groups of $n \times n$ matrices, respectively). Such admissible representations π can be written as a restricted tensor product

$$\pi = \bigotimes_v \pi_v,$$

where π_v is an irreducible representation of $\mathrm{GL}_n(K_v)$. By the work of Harish-Chandra (1923–1983), the structure of these representations is well known. Indeed, if B is the *Borel subgroup* of $\mathrm{GL}_n(K_v)$ consisting of upper triangular matrices

$$b = \begin{pmatrix} b_1 & * & \cdots & * \\ & b_2 & \cdots & * \\ & & \ddots & \\ & & & b_n \end{pmatrix},$$

for any n-tuple, $z = (z_1, z_2, \ldots, z_n) \in \mathbb{C}^n$, we define a character χ_z of B by setting

$$\chi_z(b) = |b_1|_v^{z_1} \cdots |b_n|_v^{z_n},$$

where $|\cdot|_v$ denotes the v-adic metric. We let $\widetilde{\pi}_{v,z}$ be the representation of $\mathrm{GL}_n(K_v)$ obtained by inducing χ_z from B to $\mathrm{GL}_n(K_v)$. We assume

$$\mathrm{Re}(z_1) \geq \mathrm{Re}(z_2) \geq \cdots \geq \mathrm{Re}(z_n).$$

A special case of the Langlands classification theorem shows that $\widetilde{\pi}_{v,z}$ has a unique irreducible quotient $\pi_{v,z}$. One shows that π_v is equivalent to $\pi_{v,z}$ for some z. Using this data, we define the matrix

$$A_v = \mathrm{diag}(\mathrm{N}\, v^{-z_1}, \ldots, \mathrm{N}\, v^{-z_n}),$$

where $\mathrm{N}\, v$ denotes the norm of v. For a given representation π, there is a finite set S of places such that π_v is of the type described above for $v \notin S$. For such v, we set

$$L_v(s, \pi) = \det(I - A_v\, \mathrm{N}\, v^{-s})^{-1}$$

and set

$$L_S(s, \pi) = \prod_{v \notin S} L_v(s, \pi).$$

Using known estimates of the z_i, one can show that this defines an analytic function in some fixed half-plane. Langlands obtained the meromorphic continuation of $L_S(s,\pi)$. It is possible to define $L_v(s,\pi_v)$ for $v \in S$ such that the complete L-function

$$L(s,\pi) = \prod_v L_v(s,\pi_v)$$

has a meromorphic continuation and functional equation. If π is cuspidal, then $L(s,\pi)$ extends to an entire function unless $n = 1$ and π is of the form $|\cdot|^t$ for some $t \in \mathbb{C}$. This is a celebrated theorem of Roger Godement (1921–2016) and Hervé Jacquet (born 1939). The *Ramanujan conjecture* is the assertion that the eigenvalues of A_v are of absolute value 1 for cuspidal automorphic representations. This conjecture has been proved in some special cases but remains largely an open problem.

An inspired account of the development of the Langlands program can be found in the recent autobiography of Edward Frenkel [**Fr**]. A more advanced introduction can be found in the expository article of Gelbart [**Ge**].

7.3. A Brief Overview of Elliptic Curves and Their L-functions

This is an extremely brief synopsis of a rich theory worthy of careful study. We can recommend two accessible texts for the undergraduate student. The first is by Silverman and Tate [**ST**] and the other is Chapter 18 of the book by Ireland and Rosen [**IR**]. For the ambitious student, we suggest the book by Silverman [**Sil**].

An *elliptic curve* E over a field K is given by an equation

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6$$

with $a_i \in K$, and the *discriminant* of the equation is nonzero together with a fixed basepoint. If the field K has characteristic not equal to 2 or 3, then we can describe E using a Weierstrass equation

$$y^2 = x^3 + ax + b \tag{7.1}$$

with $a, b \in K$, and the discriminant $\Delta = -16(4a^3 + 27b^2) \neq 0$.

If K is an algebraic number field, one can associate the *conductor* of E by

$$N = \prod_{\mathfrak{p}} \mathfrak{p}^{\delta(\mathfrak{p})},$$

where the product is over prime divisors of the discriminant and $\delta(\mathfrak{p})$ is a positive integer determined by the behaviour of E at the prime ideal $\mathfrak{p}$.

We say E has *good reduction* at $\mathfrak{p}$ if $\delta(\mathfrak{p}) = 0$. For such primes, the number of points modulo $\mathfrak{p}$ can be written as

$$\mathrm{N}(\mathfrak{p}) + 1 - a(\mathfrak{p})$$

with $\mathrm{N}(\mathfrak{p})$ being the norm of the ideal $\mathfrak{p}$ and

$$|a(\mathfrak{p})| \leq 2\,\mathrm{N}(\mathfrak{p})^{\frac{1}{2}}. \tag{7.2}$$

The *L-series* of E/K is defined as

$$L(E/K, s) := \prod_{\mathfrak{p}} L_{\mathfrak{p}}(E/K, s),$$

where the product is over all prime ideals $\mathfrak{p}$ of $\mathcal{O}_K$, and for $\mathfrak{p}$ coprime to N

$$L_{\mathfrak{p}}(E/K, s) = \left(1 - a(\mathfrak{p})\,\mathrm{N}(\mathfrak{p})^{-s} + \mathrm{N}(\mathfrak{p})^{1-2s}\right)^{-1}.$$

For $\mathfrak{p}|N$, one can define $L_{\mathfrak{p}}(E/K, s)$ based on its reduction modulo $\mathfrak{p}$. The bound given in (7.2) shows that $L(E/K, s)$ converges absolutely for $\mathrm{Re}(s) > 3/2$. The *Taniyama–Shimura conjecture* is the assertion that $L(E/K, s)$ is equal to $L(s - 1/2, \pi)$ for some automorphic representation π of $\mathrm{GL}_2(\mathbb{A}_K)$. When K is the rational number field, this is the celebrated theorem of Andrew Wiles (born 1953) and Breuil, Conrad, Diamond, and Taylor. It is also known for real quadratic fields. There are some partial results in the general case. In the literature, this conjecture is sometimes referred to as the *modularity conjecture*. We say E is *automorphic* over K when the conjecture holds. The work of Wiles combined with earlier work of Ken Ribet (born 1948) led to the solution of Fermat's last theorem.

The set of solutions of (7.1) with $x, y \in K$, denoted $E(K)$, can be given the structure of an abelian group. A celebrated theorem of Louis J. Mordell (1888–1972) and André Weil (1906–1998) proves

that $E(K)$ is a finitely generated abelian group when K is an algebraic number field. The number of generators is called the *rank*. That is,

$$E(K) \simeq E(K)_{\text{tors}} \oplus \mathbb{Z}^r,$$

where the torsion subgroup $E(K)_{\text{tors}}$ is finite.

The famous *Birch and Swinnerton-Dyer conjecture* predicts that $L(E/K, s)$ has a zero of order r at $s = 1$. The L-function satisfies a functional equation relating its value at s to its value at $2 - s$. Thus, the line $\text{Re}(s) = 1$ is the critical line of symmetry. There is a substantially weaker conjecture called the *parity conjecture*. This predicts that the parity of the rank equals the parity of the order of zero at $s = 1$.

These conjectures are partially motivated by a powerful analogy between $E(K)$ and $\mathcal{O}_K^\times$, the group of units of the ring of integers of $\mathcal{O}_K$. The reader will recall that $\mathcal{O}_K^\times$ is a finitely generated abelian group, and it is known that the rank of $\mathcal{O}_K^\times$ is equal to the order of the zero of $\zeta_K(s)$ at $s = 0$. This analogy has led to further conjectures about orders of zeros of L-functions attached to algebraic varieties.

One can associate to each elliptic curve E over an algebraic number field K, the *Shafarevich–Tate group*, denoted $Ш(E/K)$, which measures the deviation of the local-global principle for E. More precisely, we say two elliptic curves E_1 and E_2 over K are *locally isomorphic* if they are isomorphic (as varieties) over all completions K_v. If two curves are locally isomorphic, then Ernst S. Selmer (1920–2006) showed by giving explicit examples that it does not follow necessarily that they are globally isomorphic. For a given curve E/K, we may consider the set S of all elliptic curves E/K which are locally isomorphic to E/K. The subset of S on which E acts simply transitively can be given the structure of an abelian group, and this is $Ш(E/K)$. It is conjectured that $Ш(E/K)$ is finite, and Barry Mazur (born 1937) has shown that this conjecture is equivalent to the finiteness of S. We refer the reader to the excellent expository article of Mazur [**Maz**].

It is known that if $Ш(E/K)$ is finite, then its order must be a perfect square. Some progress on the finiteness conjecture has been made. In 1987, Karl Rubin proved this for *some* elliptic curves of rank at most one and with complex multiplication. In 1988 Victor

Kolyvagin showed the same for modular elliptic curves using an analytic theorem of Murty and Murty [**MM1**] and Bump, Friedberg, and Hoffstein [**BFH**].

7.4. Nonvanishing of L-functions and Hilbert's Problem

We give in this section a brief survey of the contents of the paper [**MP**] which connects the theory of L-functions and Hilbert's tenth problem over number fields.

Let K be an algebraic number field, and let $\mathcal{O}_K$ be its ring of integers. If we can show that $\mathbb{Z}$ has a Diophantine definition in $\mathcal{O}_K$, then Hilbert's tenth problem over $\mathcal{O}_K$ reduces the problem to Hilbert's tenth problem over $\mathbb{Z}$, which by Davis, Matiyasevič, Putnam, and Robinson, is unsolvable. This has been the approach adopted by many who have been working on this problem. Using this method, we know that Hilbert's tenth problem is unsolvable if K is totally real or is a quadratic extension of a totally real field through the work of Denef and Lipshitz (see [**D1**] and [**DL**]), and if K has exactly one nonreal Archimedean place by the work of Pheidas [**Ph**], Shlapentokh [**Sh1**], Videla [**Vi**] (independently).

To discuss briefly these developments without being stifled by technical verbiage, it is convenient to make the following definitions. A set $S \subseteq \mathcal{O}_K^n$ is said to be *Diophantine* in $\mathcal{O}_K$ if there is a polynomial

$$f \in \mathcal{O}_K[x_1, \ldots, x_n, y_1, \ldots, y_m]$$

such that

$$S = \{(a_1, \ldots, a_n) \in \mathcal{O}_K^n : \exists (b_1, \ldots, b_m) \text{ with } f(a_1, \ldots, a_n, b_1, \ldots, b_m) = 0\}.$$

It is easy to see that if $\mathbb{Z}$ is Diophantine in $\mathcal{O}_K$, then any algorithm requested by Hilbert's tenth problem for $\mathcal{O}_K$ can be modified to get a general algorithm for Hilbert's tenth problem over $\mathbb{Z}$, which is not possible. So, all methods aimed at showing that Hilbert's tenth problem is unsolvable over $\mathcal{O}_K$ have focused on showing that $\mathbb{Z}$ is Diophantine in $\mathcal{O}_K$.

More generally, we say that a number field extension L/K is *integrally Diophantine* if $\mathcal{O}_K$ is Diophantine in $\mathcal{O}_L$. This property has some nice functorial properties. For instance, if L/K is integrally Diophantine and Hilbert's tenth problem is unsolvable in $\mathcal{O}_K$, then it is unsolvable over $\mathcal{O}_L$ as well. Then there is the *transitive* property: if L/K and K/M are both integrally Diophantine, then so is L/M. If L/M is integrally Diophantine and $M \subseteq K \subseteq L$ is an intermediate field, then K/M is integrally Diophantine. Finally, if L/K_1 and L/K_2 are both integrally Diophantine, then so is $L/K_1 \cap K_2$. These results can be found in [**SS**]

These properties will be useful in our investigation of Hilbert's tenth problem in number fields. For example, we can use them to show the desired unsolvability for any *abelian* extension of $\mathbb{Q}$. Indeed, Hilbert's tenth problem is unsolvable for any cyclotomic field (which is a CM field and hence a quadratic extension of a totally real field). By the celebrated Kronecker–Weber theorem, every abelian extension of $\mathbb{Q}$ is contained in some cyclotomic field. Since it has been shown that $\mathbb{Z}$ is Diophantine for any cyclotomic field, it follows that any abelian extension is integrally Diophantine. So the result follows from the first property.

In 2002 Bjorn Poonen [**Po**] linked the theory of elliptic curves with Hilbert's tenth problem by proving the following remarkable theorem. Let L/K be a finite extension of algebraic number fields. Suppose there is an elliptic curve E/K such that

$$\text{rk } E(L) = \text{rk } E(K) = 1.$$

Then L/K is integrally Diophantine. This result motivated Poonen to ask whether such a curve exists. If we believe the Birch and Swinnerton-Dyer conjecture, the question asks for an elliptic curve E whose associated L-function has a simple zero at $s = 1$ and whose *base change* to the extension L/K also has a simple zero at $s = 1$. Since we know that the base change L-function factors as a product of the L-function of E over K and *twists* of this L-function, the condition is equivalent to the nonvanishing of all the twists.

Since our goal is to show that Hilbert's tenth problem is unsolvable for any number field $K/\mathbb{Q}$, it suffices to verify Poonen's elliptic

curve criterion for cyclic extensions of prime degree. Mazur and Rubin [**MR**] have shown that assuming that the 2-torsion part of the Shafarevich–Tate groups of elliptic curves is a perfect square (the groups are conjecturally finite; if they are finite, a famous theorem of J. W. S. Cassels implies it must be a perfect square), then Hilbert's tenth problem is unsolvable for *all* algebraic number fields.

Poonen's criterion was generalized by Cornelissen, Pheidas, and Zahidi [**CPZ**] and later by Alexandra Shlapentokh [**Sh2**] as follows. Suppose there is an elliptic curve E/K such that

$$\text{rk } E(L) = \text{rk } E(K) > 0.$$

Then L/K is integrally Diophantine. This criterion is less restricted but still amounts to showing nonvanishing of certain L-functions, assuming the Birch and Swinnerton-Dyer conjecture.

This weakened criterion allowed Murty and Pasten [**MP**] to show the following. Suppose that elliptic curves over number fields are automorphic and that they satisfy the parity conjecture. In addition, assume that for every grossencharacter η,

$$\text{ord}_{s=1} L(E/K, s, \eta) = 0 \;\Rightarrow\; \dim(E(L) \otimes \mathbb{C})^{\eta} = 0.$$

(This is often referred to as the analytic rank 0 part of the twisted Birch and Swinnerton-Dyer conjecture.) Then, for every number field K, Hilbert's tenth problem for $\mathcal{O}_K$ is unsolvable.

The analytic rank condition described above has been the focus of considerable attention ever since the work of Kolyvagin appeared. The works of Murty and Murty [**MM1**] and Bump, Friedberg, and Hoffstein [**BFH**] can be applied to attain the desired end. These ideas are amplified in a recent paper of Murty and Pasten [**MP**]. The elaborate details are beyond the scope of this monograph.

Exercises

7.1. Show that $\sqrt{2}/3$ is an algebraic number but not an algebraic integer.

7.2. Show that the ring of integers of $\mathbb{Q}(\sqrt{-1})$ is the Gaussian ring $\mathbb{Z}[\sqrt{-1}]$.

7.3. Show that at least one of the numbers $\pi + e$ and πe is transcendental. (It is conjectured that both are transcendental.)

7.4. Let (A, B) be a pair of sets of rational numbers with $B = \mathbb{Q}\backslash A$. Then (A, B) is a *Dedekind cut* if the following hold:

- $A \neq \varnothing$ and $A \neq \mathbb{Q}$;
- if $p \in A$ and $q < p$, then $q \in A$;
- if $p \in A$, then there exists $r \in A$ with $r > p$.
  - Suppose (A, B) and (C, D) are Dedekind cuts. Show that $A \subseteq C$ or $C \subseteq A$.
  - Let $r \in \mathbb{Q}$. Let $A = \{q \in \mathbb{Q} : p < r\}$ and $B = \mathbb{Q}\backslash A$. Show that (A, B) is a Dedekind cut.
  - For each natural number n, let

$$A_n = \Big\{x \in \mathbb{Q} : x < \sum_{j=0}^{n} \frac{1}{j!}\Big\}.$$

    Let $A = \bigcup_n A_n$. Show that Euler's number e is defined by the Dedekind cut (A, B) where $B = \mathbb{Q}\backslash A$.

Appendix A

Background Material

In this text we have assumed that the reader has taken some sort of advanced level math course, at least enough to be acquainted with the basics of logic, sets, relations, and functions. An introductory course to mathematical proofs, commonly offered by many universities and colleges, should be sufficient. Even if a student has not taken such a course, much of this information would be covered by a good first year course in calculus or linear algebra. In this appendix, we give a brief overview of some of this material.

For propositions P and Q, which take on either the value true, denoted T, or false, denoted F, we define the *logical connectives* $\neg$ (negation, read as "not"), $\vee$ (disjunction, read as "or"), $\wedge$ (conjunction, read as "and"), $\Longrightarrow$ (conditional or implication, read as "if, then"), and $\Longleftrightarrow$ (biconditional, read as "if and only if") in the usual way. We give their truth tables below, which show how the truth values for each connective depend on the truth of P and Q.

P	$\neg P$
T	F
F	T

P	Q	$P \vee Q$
T	T	T
T	F	T
F	T	T
F	F	F

P	Q	$P \wedge Q$
T	T	T
T	F	F
F	T	F
F	F	F

P	Q	$P \Longrightarrow Q$
T	T	T
T	F	F
F	T	T
F	F	T

P	Q	$P \Longleftrightarrow Q$
T	T	T
T	F	F
F	T	F
F	F	T

We use two quantifiers: the *universal quantifier* $\forall$ (read as "for all" or "for every") and the *existential quantifier* $\exists$ (read as "there exists" or "for some"). The proposition $\forall x P(x)$ is true if $P(x)$ is true for every value of x and is false otherwise.[1] The proposition $\exists x P(x)$ is true if there is at least one x for which $P(x)$ is true and is false otherwise. A variable that is under the scope of a quantifier is called *bound*. If a variable is not bound, then it is *free*. For example, in

$$P(x, y) \wedge \forall x Q(x, y),$$

the second appearance of the variable x is bound by the universal quantifier, while the first appearance of x and the variable y are free.

A *set* is a collection of objects. Often, sets are written in the form $\{x \in S : P(x)\}$, where S is a previously given set and $P(x)$ is a property of x. If S is implicitly clear, it may not be explicitly stated. For example, working over $\mathbb{N}$,

$$E = \{x : \exists y(x = 2y)\}$$

(here $x, y \in \mathbb{N}$) is the set of even natural numbers. If for each $x \in A$ we have $x \in B$, then we write $A \subseteq B$ and say A is a *subset* of B. For example, for the set E given above, we have $E \subseteq \mathbb{N}$.

We may build new sets from given sets A and B. The *union* of A with B is

$$A \cup B = \{x : x \in A \vee x \in B\}.$$

The *intersection* of A and B is

$$A \cup B = \{x : x \in A \wedge x \in B\}.$$

The *complement* of A is

$$A^c = \{x : x \notin A\}.$$

[1] If not otherwise stated, it is implicitly understood that we are quantifying over all x in some universe; perhaps over all natural numbers, or over all sets, for example.

Here, we understand that x are taken from some implicit universal set if not otherwise stated. For example, the complement of the set E of even natural numbers is the set of odd natural numbers, and it does not include, for example, $\sqrt{2}$ even though $\sqrt{2} \notin E$.

Given sets A and B, a *relation* from A to B is any subset of

$$\{(a, b) : a \in A \wedge b \in B\}.$$

That is, a relation R is a set consisting of some ordered pairs whose first components are elements of A and second components are elements of B. A relation from A to A is called a relation on A. If $(a, b) \in R$, then we say that a is related to b by R. For example, working over $\mathbb{N}$ we can define

$$R = \{(a, b) : \exists c(a + c = b)\}$$

(here $a, b, c \in \mathbb{N}$). This defines the usual $\leq$ relation: $(a, b) \in R$ if and only if $a \leq b$. This same definition of a relation can be extended from ordered pairs to ordered n-tuples.

Given a relation R from A to B, the *inverse relation* R^{-1} is the relation from B to A:

$$R^{-1} = \{(b, a) : (a, b) \in R\}.$$

For example, the inverse relation of $\leq$ on $\mathbb{N}$ is $\geq$.

A relation R on A is *reflexive* if for each $a \in A$, we have $(a, a) \in R$. That is, a relation is reflexive if every element of A is related to itself. R is *symmetric* if $(a, b) \in R$ implies $(b, a) \in R$. Finally, R is *transitive* if $(a, b) \in R$ and $(b, c) \in R$ implies $(a, c) \in R$. A relation that is reflexive, symmetric, and transitive is called an *equivalence relation*. If R is an equivalence relation on A and $a \in A$, then we define the *equivalence class for* a to be the set

$$\overline{a} = \{b \in A : (a, b) \in R\}.$$

That is, the equivalence class for a consists of all elements related to a. Since R is reflexive, the equivalence class for a contains at least a, and is thus nonempty. It is a fundamental theorem on equivalence relations that the distinct equivalence classes partition A into disjoint sets.

Modular arithmetic is an important application of equivalence relations. Fix a natural number $n \geq 2$. We define a relation on $\mathbb{Z}$ as follows: a is related to b if n divides $a - b$. This is an equivalence relation. We write $a \equiv b \pmod{n}$ when a is related to b. Let $a, b \in \mathbb{Z}$ have equivalence classes $\overline{a}$ and $\overline{b}$, respectively. Note that each of these are sets of integers, and if $c \in \overline{a}$, then $\overline{c} = \overline{a}$. Thus, the same equivalence class will have many different representatives. We define $\overline{a} + \overline{b}$ to be $\overline{a + b}$ and $\overline{a} \cdot \overline{b}$ to be $\overline{ab}$. It is an important exercise to show that these operations are well defined; that is, the sum and product of equivalence classes do not depend on the representatives used for each class. These operations give us *modular arithmetic*.

Given sets A and B, a *function* from A to B is a relation from A to B with the following property: for each $a \in A$, there is exactly one $b \in B$ such that a is related to b. If f is a function from A to B, we write $f : A \to B$. Since a relation is a set of ordered pairs, so is a function. A function is a relation where every element of a appears as the first coordinate in exactly one ordered pair in the relation. If f is a function and $(a, b) \in f$, we write $f(a) = b$. The use of equality here is justified by the fact that for each a there is a unique b with $(a, b) \in f$.

The set A is called the *domain* of the function, and B is called the *codomain*. The subset of the codomain defined by

$$\{b \in B : \exists a \in A(f(a) = b)\}$$

is called the *range* of f.

Let f be a function from A to B. If for all $a, b \in A$, $f(a) = f(b)$ implies $a = b$, then we say that f is *injective* or *one-to-one*. For an injective function, the elements of its codomain are mapped to at most once. If for each $b \in B$, there is some $a \in A$ with $f(a) = b$, then we say that f is *surjective* or *onto*. For a surjective function, the elements of its codomain are mapped to at least once. If f is both injective and surjective, we say that it is *bijective*. For a bijective function, each element of its codomain is mapped to exactly once. That is, a bijective function $f : A \to B$ yields a perfect pairing between the elements of A and B.

If $f : A \to B$ is bijective, then the inverse relation f^{-1} is a function from B to A. This is a fundamental result on bijective functions. When f is bijective, f is said to be *invertible*, and $f^{-1} : B \to A$ is called the *inverse* of f. If $f(a) = b$, then $f^{-1}(b) = a$.

Given functions $f : A \to B$ and $g : B \to C$, the *composition* $g \circ f$ is the function from A to C defined by $(g \circ f)(x) = g(f(x))$ for all $x \in A$. If f and g are injective, then so if $g \circ f$. If f and g are surjective, then so is $g \circ f$. Thus the composition of two bijective functions is bijective. If $f : A \to B$ is bijective and, hence, invertible, then $f^{-1} \circ f = i_A$, where $i_A : A \to A$ is the *identity function on* A defined by $i_A(a) = a$ for all $a \in A$.

Bibliography

[AMS] F. E. Browder (ed.), *Mathematical developments arising from Hilbert problems*, Proceedings of Symposia in Pure Mathematics, Vol. XXVIII, American Mathematical Society, Providence, R. I., 1976. MR0419125

[Ba1] A. Baker, *A concise introduction to the theory of numbers*, Cambridge University Press, Cambridge, 1984. MR781734

[Ba2] A. Baker, *A comprehensive course in number theory*, Cambridge University Press, Cambridge, 2012. MR2954465

[BFH] D. Bump, S. Friedberg, and J. Hoffstein, *On some applications of automorphic forms to number theory*, Bull. Amer. Math. Soc. (N.S.) **33** (1996), no. 2, 157–175, DOI 10.1090/S0273-0979-96-00654-4. MR1359575

[CPZ] G. Cornelissen, T. Pheidas, and K. Zahidi, *Division-ample sets and the Diophantine problem for rings of integers* (English, with English and French summaries), J. Théor. Nombres Bordeaux **17** (2005), no. 3, 727–735. MR2212121

[Cr] J. N. Crossley, C. J. Ash, C. J. Brickhill, J. C. Stillwell, and N. H. Williams, *What is mathematical logic?*, Oxford University Press, London-New York, 1972. Oxford Paperbacks University Series, No. 60. MR0414308

[Dav] H. Davenport, *Multiplicative number theory*, 2nd ed., Graduate Texts in Mathematics, vol. 74, Springer-Verlag, New York-Berlin, 1980. Revised by Hugh L. Montgomery. MR606931

[Da1] M. Davis, *Computability and unsolvability*, McGraw-Hill Series in Information Processing and Computers, McGraw-Hill Book Co., Inc., New York-Toronto-London, 1958. MR0124208

[Da2] M. Davis, *Hilbert's tenth problem is unsolvable*, Amer. Math. Monthly **80** (1973), 233–269, DOI 10.2307/2318447. MR0317916

[Da3] M. Davis, *On the number of solutions of Diophantine equations*, Proc. Amer. Math. Soc. **35** (1972), 552–554, DOI 10.2307/2037646. MR0304347

[DMR] M. Davis, Y. Matijasevič, and J. Robinson, *Hilbert's tenth problem: Diophantine equations: positive aspects of a negative solution*, Mathematical developments arising from Hilbert problems (Proc. Sympos. Pure Math., Vol. XXVIII, Northern Illinois Univ., De Kalb, Ill., 1974), Amer. Math. Soc., Providence, R. I., 1976, pp. 323–378. MR0432534

[Daw] J. W. Dawson Jr., *Logical dilemmas*: *The life and work of Kurt Gödel*, A K Peters, Ltd., Wellesley, MA, 1997. MR1429389

[De] R. Dedekind, *Theory of algebraic integers*, Cambridge Mathematical Library, Cambridge University Press, Cambridge, 1996. Translated from the 1877 French original and with an introduction by John Stillwell. MR1417492

[D] J. Denef, *Hilbert's tenth problem for quadratic rings*, Proc. Amer. Math. Soc. **48** (1975), 214–220, DOI 10.2307/2040720. MR0360513

[D1] J. Denef, *Diophantine sets over algebraic integer rings. II*, Trans. Amer. Math. Soc. **257** (1980), no. 1, 227–236, DOI 10.2307/1998133. MR549163

[DL] J. Denef and L. Lipshitz, *Diophantine sets over some rings of algebraic integers*, J. London Math. Soc. (2) **18** (1978), no. 3, 385–391, DOI 10.1112/jlms/s2-18.3.385. MR518221

[DPR] M. Davis, H. Putnam, and J. Robinson, *The decision problem for exponential diophantine equations*, Ann. of Math. (2) **74** (1961), 425–436, DOI 10.2307/1970289. MR0133227

[Du] A. K. Dutta, *Kuṭṭaka, bhāvanā and cakravāla*, Studies in the history of Indian mathematics, Cult. Hist. Math., vol. 5, Hindustan Book Agency, New Delhi, 2010, pp. 145–199. MR2648498

[En] H. B. Enderton, *A mathematical introduction to logic*, 2nd ed., Harcourt/Academic Press, Burlington, MA, 2001. MR1801397

[Fo] B. Fodden, *Diophantine equations and the generalized Riemann hypothesis*, J. Number Theory **131** (2011), no. 9, 1672–1690, DOI 10.1016/j.jnt.2011.01.017. MR2802141

[Fr] E. Frenkel, *Love and math*: *The heart of hidden reality*, Basic Books, New York, 2013. MR3155773

[Ge] S. Gelbart, *An elementary introduction to the Langlands program*, Bull. Amer. Math. Soc. (N.S.) **10** (1984), no. 2, 177–219, DOI 10.1090/S0273-0979-1984-15237-6. MR733692

[Go] D.C. Goldrei, *Classic set theory for guided independent study*, CRC Press, Boca Raton, FL, 1998.

[Gra] J. J. Gray, *The Hilbert challenge*, Oxford University Press, Oxford, 2000. MR1828558

[GT] B. Green and T. Tao, *The primes contain arbitrarily long arithmetic progressions*, Ann. of Math. (2) **167** (2008), no. 2, 481–547, DOI 10.4007/annals.2008.167.481. MR2415379

[Ha] P. R. Halmos, *Naive set theory*, Springer-Verlag, New York-Heidelberg, 1974. Reprint of the 1960 edition; Undergraduate Texts in Mathematics. MR0453532

[HW] G. H. Hardy and E. M. Wright, *An introduction to the theory of numbers*, 6th ed., Oxford University Press, Oxford, 2008. Revised by D. R. Heath-Brown and J. H. Silverman; With a foreword by Andrew Wiles. MR2445243

[Haw] S. Hawking (ed.), *God created the integers*: *The mathematical breakthroughs that changed history*, Running Press, Philadelphia, PA, 2007. Edited, with commentary, by Stephen Hawking. MR2382242

[Hil02] D. Hilbert, *Mathematical Problems*, Bull. Amer. Math. Soc. **8** (1902), no. 10, 437–479. MR1557926

[Ho] R. Hodel, *An Introduction to Mathematical Logic*, Dover Publications, New York, 2013.

[Hod] A. Hodges, *Alan Turing: the enigma*, The centenary edition, Princeton University Press, Princeton, NJ, 2012. With a foreword by Douglas Hofstadter and a new preface by the author. MR2963548

[HR] P. Howard and J. E. Rubin, *Consequences of the axiom of choice*, Mathematical Surveys and Monographs, vol. 59, American Mathematical Society, Providence, RI, 1998. With 1 IBM-PC floppy disk (3.5 inch; WD). MR1637107

[IR] K. F. Ireland and M. I. Rosen, *A classical introduction to modern number theory*, Graduate Texts in Mathematics, vol. 84, Springer-Verlag, New York-Berlin, 1982. Revised edition of *Elements of number theory*. MR661047

[Je] T. Jech, *Set theory*: *The third millennium edition*, Springer Monographs in Mathematics, Springer-Verlag, Berlin, 2003. revised and expanded. MR1940513

[Jo1] J. P. Jones, *Three universal representations of recursively enumerable sets*, J. Symbolic Logic **43** (1978), no. 2, 335–351, DOI 10.2307/2272832. MR0498049

[Jo2] J. P. Jones, *Universal Diophantine equation*, J. Symbolic Logic **47** (1982), no. 3, 549–571, DOI 10.2307/2273588. MR666816

[JSWW] J. P. Jones, D. Sato, H. Wada, and D. Wiens, *Diophantine representation of the set of prime numbers*, Amer. Math. Monthly **83** (1976), no. 6, 449–464, DOI 10.2307/2318339. MR0414514

[KP] L. Kirby and J. Paris, *Accessible independence results for Peano arithmetic*, Bull. London Math. Soc. **14** (1982), no. 4, 285–293, DOI 10.1112/blms/14.4.285. MR663480

[Ku] K. Kunen, *Set theory*, Studies in Logic (London), vol. 34, College Publications, London, 2011. MR2905394

[LK] C. Leary, and L. Kristiansen, *A friendly introduction to mathematical logic*, Second edition, Milne Library, Genesco, NY, 2015.

[Ma] Y. V. Matiyasevič, *Hilbert's tenth problem*, Foundations of Computing Series, MIT Press, Cambridge, MA, 1993. Translated from the 1993 Russian original by the author; With a foreword by Martin Davis. MR1244324

[Maz] B. Mazur, *On the passage from local to global in number theory*, Bull. Amer. Math. Soc. (N.S.) **29** (1993), no. 1, 14–50, DOI 10.1090/S0273-0979-1993-00414-2. MR1202293

[MR] B. Mazur and K. Rubin, *Ranks of twists of elliptic curves and Hilbert's tenth problem*, Invent. Math. **181** (2010), no. 3, 541–575, DOI 10.1007/s00222-010-0252-0. MR2660452

[Mu] M. R. Murty, *Problems in analytic number theory*, 2nd ed., Graduate Texts in Mathematics, vol. 206, Springer, New York, 2008. Readings in Mathematics. MR2376618

[ME] M. R. Murty and J. Esmonde, *Problems in algebraic number theory*, 2nd ed., Graduate Texts in Mathematics, vol. 190, Springer-Verlag, New York, 2005. MR2090972

[MM1] M. R. Murty and V. K. Murty, *Mean values of derivatives of modular L-series*, Ann. of Math. (2) **133** (1991), no. 3, 447–475, DOI 10.2307/2944316. MR1109350

[MM2] M. R. Murty and V. K. Murty, *Non-vanishing of L-functions and applications*, Progress in Mathematics, vol. 157, Birkhäuser Verlag, Basel, 1997. MR1482805

[MP] M. R. Murty and H. Pasten, *Elliptic curves, L-functions, and Hilbert's tenth problem*, J. Number Theory **182** (2018), 1–18, DOI 10.1016/j.jnt.2017.07.008. MR3703929

[MR] M. R. Murty and P. Rath, *Transcendental numbers*, Springer, New York, 2014. MR3134556

[Ph] T. Pheidas, *Hilbert's tenth problem for a class of rings of algebraic integers*, Proc. Amer. Math. Soc. **104** (1988), no. 2, 611–620, DOI 10.2307/2047021. MR962837

[Po] B. Poonen, *Using elliptic curves of rank one towards the undecidability of Hilbert's tenth problem over rings of algebraic integers*, Algorithmic number theory (Sydney, 2002), Lecture Notes in Comput. Sci., vol. 2369, Springer, Berlin, 2002, pp. 33–42, DOI 10.1007/3-540-45455-1_4. MR2041072

[SS] H. N. Shapiro and A. Shlapentokh, *Diophantine relationships between algebraic number fields*, Comm. Pure Appl. Math. **42** (1989), no. 8, 1113–1122, DOI 10.1002/cpa.3160420805. MR1029120

[Sh1] A. Shlapentokh, *Extension of Hilbert's tenth problem to some algebraic number fields*, Comm. Pure Appl. Math. **42** (1989), no. 7, 939–962, DOI 10.1002/cpa.3160420703. MR1008797

[Sh2] A. Shlapentokh, *Elliptic curves retaining their rank in finite extensions and Hilbert's tenth problem for rings of algebraic numbers*, Trans. Amer. Math. Soc. **360** (2008), no. 7, 3541–3555, DOI 10.1090/S0002-9947-08-04302-X. MR2386235

[Sh3] A. Shlapentokh, *Hilbert's tenth problem: Diophantine classes and extensions to global fields*, New Mathematical Monographs, vol. 7, Cambridge University Press, Cambridge, 2007. MR2297245

[Sie] C. L. Siegel, *Zur Theorie der quadratischen Formen* (German), Nachr. Akad. Wiss. Göttingen Math.-Phys. Kl. II (1972), 21–46. MR0311578

[Sil] J. H. Silverman, *The arithmetic of elliptic curves*, Graduate Texts in Mathematics, vol. 106, Springer-Verlag, New York, 1986. MR817210

[ST] J. H. Silverman and J. Tate, *Rational points on elliptic curves*, Undergraduate Texts in Mathematics, Springer-Verlag, New York, 1992. MR1171452

[Sm] C. Smoryński, *Logical number theory I*, Springer-Verlag, Berlin, 1991.

[Vi] C. Videla, *On Hilbert's tenth problem.* Atas da Xa Escola de Algebra, Vitoria, ES, Brasil Colecao Atas **16** Sociedade Brasileira de Matematica (1989), 95–108.

[We] A. Weil, *Number theory: An approach through history; From Hammurapi to Legendre*, Birkhäuser Boston, Inc., Boston, MA, 1984. MR734177

[Yan] B. H. Yandell, *The honors class: Hilbert's problems and their solvers*, A K Peters, Ltd., Natick, MA, 2002. MR1880187

Index

Selected Published Titles in This Series

88 **M. Ram Murty and Brandon Fodden,** Hilbert's Tenth Problem, 2019

87 **Matthew Katz and Jan Reimann,** An Introduction to Ramsey Theory, 2018

86 **Peter Frankl and Norihide Tokushige,** Extremal Problems for Finite Sets, 2018

85 **Joel H. Shapiro,** Volterra Adventures, 2018

84 **Paul Pollack,** A Conversational Introduction to Algebraic Number Theory, 2017

83 **Thomas R. Shemanske,** Modern Cryptography and Elliptic Curves, 2017

82 **A. R. Wadsworth,** Problems in Abstract Algebra, 2017

81 **Vaughn Climenhaga and Anatole Katok,** From Groups to Geometry and Back, 2017

80 **Matt DeVos and Deborah A. Kent,** Game Theory, 2016

79 **Kristopher Tapp,** Matrix Groups for Undergraduates, Second Edition, 2016

78 **Gail S. Nelson,** A User-Friendly Introduction to Lebesgue Measure and Integration, 2015

77 **Wolfgang Kühnel,** Differential Geometry: Curves — Surfaces — Manifolds, Third Edition, 2015

76 **John Roe,** Winding Around, 2015

75 **Ida Kantor, Jiří Matoušek, and Robert Šámal,** Mathematics++, 2015

74 **Mohamed Elhamdadi and Sam Nelson,** Quandles, 2015

73 **Bruce M. Landman and Aaron Robertson,** Ramsey Theory on the Integers, Second Edition, 2014

72 **Mark Kot,** A First Course in the Calculus of Variations, 2014

71 **Joel Spencer,** Asymptopia, 2014

70 **Lasse Rempe-Gillen and Rebecca Waldecker,** Primality Testing for Beginners, 2014

69 **Mark Levi,** Classical Mechanics with Calculus of Variations and Optimal Control, 2014

68 **Samuel S. Wagstaff, Jr.,** The Joy of Factoring, 2013

67 **Emily H. Moore and Harriet S. Pollatsek,** Difference Sets, 2013

66 **Thomas Garrity, Richard Belshoff, Lynette Boos, Ryan Brown, Carl Lienert, David Murphy, Junalyn Navarra-Madsen, Pedro Poitevin, Shawn Robinson, Brian Snyder, and Caryn Werner,** Algebraic Geometry, 2013

For a complete list of titles in this series, visit the AMS Bookstore at **www.ams.org/bookstore/stmlseries/**.

Hilbert's tenth problem is one of 23 problems proposed by David Hilbert in 1900 at the International Congress of Mathematicians in Paris. These problems gave focus for the exponential development of mathematical thought over the following century. The tenth problem asked for a general algorithm to determine if a given Diophantine equation has a solution in integers. It was finally resolved in a series of papers written by Julia Robinson, Martin Davis, Hilary Putnam, and finally Yuri Matiyasevich in 1970. They showed that no such algorithm exists.

This book is an exposition of this remarkable achievement. Often, the solution to a famous problem involves formidable background. Surprisingly, the solution of Hilbert's tenth problem does not. What is needed is only some elementary number theory and rudimentary logic. In this book, the authors present the complete proof along with the romantic history that goes with it. Along the way, the reader is introduced to Cantor's transfinite numbers, axiomatic set theory, Turing machines, and Gödel's incompleteness theorems.

Copious exercises are included at the end of each chapter to guide the student gently on this ascent. For the advanced student, the final chapter highlights recent developments and suggests future directions. The book is suitable for undergraduates and graduate students. It is essentially self-contained.

ISBN 978-1-4704-4399-3

9 781470 443993

STML/88

For additional information
and updates on this book, visit
www.ams.org/bookpages/stml-88