Elementary Number Theory
&
Elliptic Curves

William Stein

December 11, 2002

Contents

1
Preface

This is a textbook about classical elementary number theory and elliptic curves. The first part discusses elementary topics such as primes, factorization, continued fractions, and quadratic forms, in the context of cryptography, computation, and deep open research problems. The second part is about elliptic curves, their applications to algorithmic problems, and their connections with problems in number theory such as Fermat's Last Theorem, the Congruent Number Problem, and the Conjecture of Birch and Swinnerton-Dyer.

The intended audience of this book is an undergraduate with some familiarity with basic abstract algebra, e.g., rings, fields, and finite abelian groups. For the second part, some prior exposure to basic complex analysis is useful but not necessary.

Our approach in the first part is classical except that we include a large number of explicit examples and teach the reader how to find more examples using a computer. We also discuss current relevant world records and open problems. Because of space constraints, the second part, about
elliptic curves, does not contain complete proofs.[1] 1

Our goal is to convey the central importance of elliptic curves in number theory and give a feeling for the big open problems about them without becoming overwhelmend by technical details.

---

[1] Or maybe it does; we shall see!

Acknowledgement. I would like to thank Lawrence Cabusora for carefully reading the first draft of this book and making many helpful comments. Brian Conrad made clarifying comments on the first 30 pages, which I've included. Noam Elkies made some comments about Section 3.2. I would also like to thank the students of my Math 124 course at Harvard during the Fall of 2001 and 2002, who provided the first audience for this book, as well as David Savitt for conversations.

1. Peter Hawthorne (discussions about algebra; helped write ...)

2. Seth Kleinerman (e; finding many typos)

I also found LaTeX, xfig, MAGMA, PARI, and Emacs to be extremely helpful in the preparation of this manuscript.

Warning.

The version of this book that you are currently looking at is not finished. Part I was created for a course based on Davenport's *The Higher Arithmetic*, so some parts might be uncomfortably similar to that book. The author intends to remove any such unintentional similarity before this book is officially published. Also, there are a few pictures (in particular, of Diffie and Hellman) that were swiped from other books without permission; this was fair use for lecture notes during a course, but not for a textbook, so this will have to be remedied.

2
Introduction

This book is divided into three parts. The first is about several standard topics in elementary number theory including primes and congruences (Chapter 3), quadratic reciprocity (Chapter 6), continued fractions (Chapter 7), and binary quadratic forms (Chapter 9), with motivation from cryptography (Chapter 4). The second is about elliptic curves and the central role they play in modern number theory. We will discuss their use in algorithmic applications (Chapter 11), their role in the proof of Fermat's Last Theorem (Chapter 12), the most central unsolved conjecture about them (Chapter 13), and their connection with the congruent number problem (Chapter **??**). The third is about how to use a computer in number theory.

For the first part of the book, some mathematical maturity and knowledge of basic abstract algebra is assumed on the part of the reader. The second part also assumes some background in analysis and a willingness to take a few statements on faith.

2.1 Elementary Number Theory

2.1.1 Prime Factorization of Integers

Remember writing integers (whole numbers) as products of primes? For example, $12 = 2 \cdot 2 \cdot 3$, as illustrated in Figure 2.1.

Does every positive integer factor as a product of primes? If so, how difficult is it to find factorizations? For example, factoring US social security

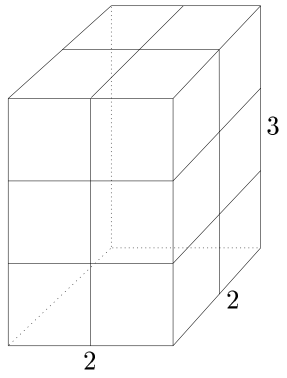


FIGURE 2.1. We have $12 = 2 \cdot 2 \cdot 3$

numbers, which have 9 digits, is easy enough that the *on*Hand wrist watch quickly does it (see [Mat]). What about bigger numbers?

These questions are important to your everyday life, because the popular RSA public-key cryptosystem relies on the difficulty of factoring large numbers quickly (see Section 4.2).

2.1.2 Congruences and Public-Key Cryptography

We say that integers a and b are congruent modulo an integer n if there is an integer k such that $a = b+nk$. That a and b are congruent means you can get from a to b by adding or subtracting copies of n. For example, $26 \equiv 2 \pmod{12}$ since $26 = 2 + 12 \cdot 2$. We will extensively study arithmetic with integers modulo n in Chapter 3. Then in Chapter 4 we will see how the RSA cryptosystem uses arithmetic with the numbers modulo n to send messages in view of an adversary without their true portent being discovered by the adversary.

2.1.3 Computers and Telescopes

A computer is to a number theorist like a telescope to an astronomer. It would be a shame to study astronomy without learning about telescopes; likewise, in Part 3 of this book you will learn how to look at the integers through the enhancing power of a computer.

2.1.4 Quadratic Reciprocity

One of the most celebrated theorems of classical number theory is Gauss's quadratic reciprocity law. One application is that it gives the following simple criterion for whether or not 5 is a square modulo an odd prime p: the number 5 is a perfect square modulo p if and only if p is congruent to 1 or 4 modulo 5. This result is impressive because it is extremely easy to be convinced that it is true by numerical observation, but difficult to prove. For more details, including the statement with 5 replaced by any odd prime, see Chapter 6.

2.1.5 Continued Fractions

A continued fraction is an expression of the form

$$a_0 + \cfrac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \cdots}}}.$$

Continued fractions have surprising applications all over number theory. They provide new insight into numbers of the form $a + b\sqrt{d}$ with a and b rational and d positive. They are useful in understanding the "modular group" $\mathrm{SL}(2, \mathbb{Z})$ of 2×2 integer matrices with determinant 1, which plays a crucial role in the theory of elliptic curves. From a computational point of view, continued fractions give rise to a powerful algorithm for recognizing a rational number x from a partial decimal expansion of x. This is frequently useful because such partial decimal expansions are often output by various algorithms. See Chapter 7 for much more.

2.1.6 Sums of Two Squares and Binary Quadratic Forms

Let n be your favorite positive integer. Is n the sum of two perfect squares? For example, 7 is not a sum of two squares, but 13 is. In Chapter 9 you will learn a beautiful criterion for whether or not a number is a sum of two squares. More generally, we will study binary quadratic forms $ax^2 + bxy + cy^2$, which provide a concrete glimpse into some of the central problems of algebraic number theory.

2.2 Elliptic Curves

An elliptic curve over $\mathbb{Q}$ is a curve of the form

$$y^2 = x^3 + ax + b,$$

where a and b are rational numbers and $x^3 + ax + b$ has distinct complex roots. The set

$$E(\mathbb{Q}) = \{(x, y) \in \mathbb{Q} \times \mathbb{Q} : y^2 = x^3 + ax + b\} \cup \{\mathcal{O}\}$$

of rational points on E is of great interest. (Here $\mathcal{O}$ is a rational point on E "at infinity".) The set $E(\mathbb{Q})$ is sometimes finite and sometimes infinite. For example, if E is defined by $y^2 = x^3 + x$ then $E(\mathbb{Q})$ is finite (see

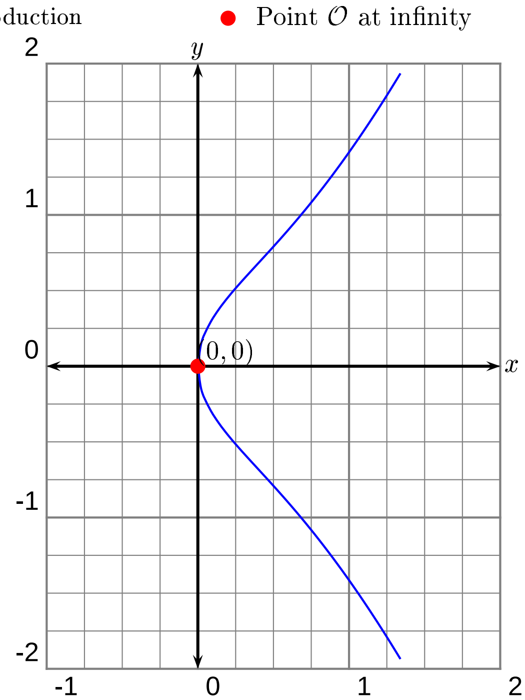


FIGURE 2.2. The Two Rational Points on the Elliptic Curve $y^2 = x^3 + x$

Figure 2.2), but if E is given by $y^2 = x^3 + 100x$, then $E(\mathbb{Q})$ is infinite. Birch and Swinnerton-Dyer gave a beautiful conjectural criterion that they believe predicts whether or not $E(\mathbb{Q})$ is infinite (see Chapter 13). To try and understand $E(\mathbb{Q})$ better, we find that this set has the additional structure of finitely generated abelian group: given two elements of $E(\mathbb{Q})$, there is a way to "add" them together to obtain another element of $E(\mathbb{Q})$ (this addition is *not* coordinate wise). Moreover, there is a finite set of elements of $E(\mathbb{Q})$ so that every element of $E(\mathbb{Q})$ is a obtained by adding together elements from this finite list. [1] 1

2.2.1 Algorithmic Applications

Elliptic curves are crucial to modern factorization methods, and elliptic curves over finite fields provide valuable alternative cryptosystems (see Chapter 11).

2.2.2 Theoretical Applications

Many exciting problems in number theory can be translated into questions about elliptic curves. For example, Fermat's Last Theorem, which asserts

[1] Draw graph of Riemann surface to clarify role of $\mathcal{O}$.

that $x^n + y^n = z^n$ has no positive integer solutions when $n > 2$, was proved by Andrew Wiles who showed that counterexamples to Fermat's Last Theorem would give rise to impossibly bizarre elliptic curves (see Chapter 12).

The ancient congruent number problem asks for an algorithm to decide whether an integer is the area of a right triangle with rational side lengths. This question is equivalent to a question about elliptic curves that has almost, but not entirely, been solved. The key mising ingredient is a proof of a certain case of the Birch and Swinnerton-Dyer conjecture (see Chapter **??**).

2.3 Notation and Conventions

We use the standard notation $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ for the rings of natural, integer, rational, real, and complex numbers, respectively. We use the words proposition, theorem, lemma, corollary, etc., in their standard mathematical way. Thus usually a proposition is a routine assertion, a theorem a deeper culmination of ideas, a lemma something that will be used later to prove a proposition or theorem, and a corollary an easy consequence of a proposition, theorem, or lemma.

Part I

Elementary Number Theory

3
Primes and Congruences

Prime numbers are the foundation from which the integers, and hence much of number theory, is built. Congruences between integers lead to the ring $\mathbb{Z}/n = \{0, 1, \ldots, n-1\}$ of equivalence classes of integers modulo n. Arithmetic in this ring is critical for every cryptosystem discussed in this book, and plays a key role in the elliptic curve factorization method (Section 11.2) and the Birch and Swinnerton-Dyer conjecture (Chapter 13).

In Section 3.1 we describe how the integers are built out of the mysterious sequence $2, 3, 5, 7, 11, \ldots$ of prime numbers. In Section 3.2 we discuss theorems about the set of primes numbers, starting with Euclid's proof that this set is infinite, then explore the distribution of primes via the prime number theorem and the Riemann Hypothesis (without proofs). Section 3.3 is about congruences modulo n and simple linear equations in the the ring $\mathbb{Z}/n$. In Section 3.4 we prove the Chinese Remainder Theorem, which describes how to solve certain systems of equations modulo n; we also use this theorem to establish the multiplicativity of the Euler φ function. Section 3.5.2 is about how being able to quickly compute huge powers in the integers modulo n leads to a way to quickly decide, with high probability, whether or not a number is prime.

3.1 Prime Factorization

3.1.1 Prime Numbers

The set of *natural numbers* is

$$\mathbb{N} = \{1, 2, 3, 4, \ldots\},$$

the set of *integers* is

$$\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}.$$

They are denoted by $\mathbb{Z}$ because the German word for the integers is **Z**ahlen, and Germans laid the foundations of number theory.

Definition 3.1.1. If $a, b \in \mathbb{Z}$ then we say that a divides b, written $a \mid b$, if $ac = b$ for some $c \in \mathbb{Z}$. We say that a does not divide b, written $a \nmid b$ if there is no $c \in \mathbb{Z}$ such that $ac = b$.

To save time, we write

$$a \mid b.$$

For example, $2 \mid 6$ and $389 \mid 97734562907$. Also, everything divides 0, and 0 divides only 0.

Definition 3.1.2. We say that a natural number $n > 1$ is *prime* if 1 and n are the only positive divisors of n, and we call n *composite* otherwise. The number 1 is neither prime nor composite.

Thus the primes are

$$2, 3, 5, 7, 11, \ldots, 389, \ldots, 2003, \ldots.$$

and the composites are

$$4, 6, 8, 9, 10, 12, \ldots, 666 = 2 \cdot 3^2 \cdot 37, \ldots, 2001 = 3 \cdot 23 \cdot 29, \ldots.$$

What about 1? One reason that we don't call 1 prime, is that Theorem 3.1.5 below asserts that every positive integer is a product of primes in a unique way; if 1 were prime, then this uniqueness would be destroyed. It is best to think of 1 as a *unit* in $\mathbb{Z}$, i.e., a number with a multiplicative inverse in $\mathbb{Z}$, and think of the natural numbers as divided into three classes: primes, composites, and units. In rings which are more complicated than $\mathbb{Z}$, this distinction is easier to appreciate (e.g., in $\mathbb{Z}[\sqrt{2}] = \{a+b\sqrt{2} : a, b \in \mathbb{Z}\}$, the element $1 + \sqrt{2}$ is a unit because $(1 + \sqrt{2})(-1 + \sqrt{2}) = 1$. For future use, we formalize the definition of unit.

Definition 3.1.3 (Unit). Let R be a ring. An element $x \in R$ is a *unit* if there exists $y \in R$ such that $xy = yx = 1$.

Remark 3.1.4. Before the influence of abstract algebra on number theory the picture was less clear. For example, in 1914 Dick Lehmer, considered 1 to be prime (see [Leh14]).

Every natural number is built, in a unique way, out of prime numbers.

Theorem 3.1.5 (Fundamental Theorem of Arithmetic). *Every positive integer can be written as a product of primes, and this expression is unique (up to order).*

This theorem, which we will prove in Section 3.1.4, is trickier to prove than you might first think. First, we are fortunate that there are any primes at all: if the natural numbers are replaced by the positive rational numbers then there are no primes; e.g., $2 = \frac{1}{2} \cdot 4$, so "$\frac{1}{2} \mid 2$" in the sense that there is a $c \in \mathbb{Q}$ such that $\frac{1}{2}c = 2$. Second, we are fortunate that factorization is unique in $\mathbb{Z}$, since there are simple rings where unique factorization fails. For example, it fails in

$$\mathbb{Z}[\sqrt{-5}] = \{a + b\sqrt{-5} : a, b \in \mathbb{Z}\},$$

where 6 factors in two different irreducible ways:

$$2 \cdot 3 = 6 = (1 + \sqrt{-5})(1 - \sqrt{-5}).$$

See Exercise 3.

3.1.2 The Greatest Common Divisor

We will use the notion of greatest common divisor of two integers to prove that if p is a prime and $p \mid ab$, then $p \mid a$ or $p \mid b$. This is the key step in our proof of Theorem 3.1.5.

Definition 3.1.6. Let $\gcd(a, b) = \max\{d : d \mid a \text{ and } d \mid b\}$, unless both a and b are 0 in which case $\gcd(0, 0) = 0$.

For example, $\gcd(1, 2) = 1$, $\gcd(3, 27) = 3$, and for any a, $\gcd(0, a) = \gcd(a, 0) = a$.

The greatest common divisor of two numbers, even large numbers, is surprisingly easy to compute. For example, let's compute $\gcd(2261, 1275)$. First, we recall the division algorithm, which you might recall from elementary school when you learned long division with remainder:

Algorithm 3.1.7 (Division Algorithm). Suppose that a and b are natural numbers. Then there exist unique nonnegative integers q and r such that $0 \le r < b$ and $a = bq + r$.

We use the division algorithm repeatedly to compute $\gcd(2261, 1275)$. Dividing 2261 by 1275 we find that

$$2261 = 1 \cdot 1275 + 986,$$

so $q = 1$ and $r = 986$. Notice that if a natural number d divides both 2261 and 1275, then d divides their difference 986 and d still divides 1275. On the other hand, if d divides both 1275 and 986, then it has got to divide their sum 2261 as well! We have made progress:

$$\gcd(2261, 1275) = \gcd(1275, 986).$$

Repeating, we have

$$1275 = 1 \cdot 986 + 289,$$

so $\gcd(1275, 986) = \gcd(986, 289)$. Keep going:

$$\begin{aligned} 986 &= 3 \cdot 289 + 119 \\ 289 &= 2 \cdot 119 + 51 \\ 119 &= 2 \cdot 51 + 17. \end{aligned}$$

Thus $\gcd(2261, 1275) = \cdots = \gcd(51, 17)$, which is 17 because $17 \mid 51$. Thus

$$\gcd(2261, 1275) = 17.$$

Aside from tedious arithmetic, that was quick and systematic. [1] 1

Algorithm 3.1.8 (Euclid's Algorithm for Computing GCDs). Fix $a, b \in \mathbb{N}$ with $a > b$. Using the division algorithm, write $a = bq + r$, with $0 \leq r < b$. Then, as above,

$$\gcd(a, b) = \gcd(b, r).$$

Let $a_1 = b$, $b_1 = r$, and repeat until the remainder is 0. Since the remainders form a decreasing sequence of nonnegative numbers, this process terminates.

Example 3.1.9. Set $a = 15$ and $b = 6$.

$$\begin{array}{rclcl} 15 &=& 6 \cdot 2 + 3 & \qquad & \gcd(15, 6) = \gcd(6, 3) \\ 6 &=& 3 \cdot 2 + 0 & \qquad & \gcd(6, 3) = \gcd(3, 0) = 3 \end{array}$$

Note that we can just as easily do an example that is ten times as big, an observation that will be important in the proof of Theorem 3.1.11 below.

Example 3.1.10. Set $a = 150$ and $b = 60$.

$$\begin{array}{rclcl} 150 &=& 60 \cdot 2 + 30 & \qquad & \gcd(150, 60) = \gcd(60, 30) \\ 60 &=& 30 \cdot 2 + 0 & \qquad & \gcd(60, 30) = \gcd(30, 0) = 30 \end{array}$$

With Euclid's algorithm in hand, we can prove that if a prime divides the product of two numbers, then it has got to divide one of them. This result is the key to proving that prime factorization is unique.

[1] Something about complexity.

Theorem 3.1.11 (Euclid). *Let p be a prime and $a, b \in \mathbb{N}$. If $p \mid ab$ then $p \mid a$ or $p \mid b$.*

The reader may think that this theorem is "intuitively obvious", but that is only because the fundamental theorem of arithmetic (Theorem 3.1.5) is deeply ingrained as a source of intuition. Yet, Theorem 3.1.11 will be needed to prove the fundamental theorem of arithmetic.

Proof. If $p \mid a$ we are done. If $p \nmid a$ then $\gcd(p, a) = 1$, since only 1 and p divide p. Stepping through Algorithm 3.1.8, an in Example 3.1.10, we see that $\gcd(pb, ab) = b$. At each step, we simply multiply the equation through by b. Since $p \mid pb$ and, by hypothesis, $p \mid ab$, it follows that

$$p \mid \gcd(pb, ab) = b.$$

□

3.1.3 Numbers Factor as Products of Primes

In this section, we prove that every natural number factors as a product of primes. Then we discuss the difficulty of finding such a decomposition in practice. We will wait until Section 3.1.4 to prove that factorization is unique.

As a first example, let $n = 1275$. Since $17 \mid 1275$, n is definitely composite, $n = 17 \cdot 75$. Next, 75 is $5 \cdot 15 = 5 \cdot 5 \cdot 3$, and we find that $1275 = 3 \cdot 5 \cdot 5 \cdot 17$. Generalizing this process proves the following proposition:

Proposition 3.1.12. *Every natural number is a product of primes.*

Proof. Let n be a natural number. If $n = 1$, then n is the empty product of primes. If n is prime, we are done. If n is composite, then $n = ab$ with $a, b < n$. By induction, a and b are products of primes, so n is also a product of primes. □

Two questions: is this factorization unique, and how quickly can we find a factorization? What if we had done something differently when breaking 1275 apart as a product of primes? Could the primes that show up be different? Let's try: we have $1275 = 5 \cdot 255$. Now $255 = 5 \cdot 51$ and $51 = 17 \cdot 3$, so the factorization is the same, as asserted by Theorem 3.1.5 above.

Regarding the second question, it is still unknown just how clever we can be at factoring.

Open Problem 3.1.13. *Is there an algorithm which can factor any integer n in polynomial time?*

By "algorithm" we mean an algorithm in the sense of classical computer science, i.e., a sequence of instructions that can be run on a classical computer, which is guaranteed to terminate. By "polynomial time" we mean that there is a polynomial $f(x)$ such that for any n the number of steps needed by the algorithm to factor n is less than $f(\log_{10}(n))$. (Note that $\log_{10}(n)$ is a good approximation for the number of digits of the input n to the algorithm.) We will discuss one of the fastest known factoring algorithms in Section 11.2.

Peter Shor [Sho97] devised a polynomial time algorithm for factoring integers on quantum computers. We will not discuss his algorithm further, except to note that IBM built a quantum computer out of a "billion-billion custom-designed molecules" in December 2001 that used Shor's algorithm to factor 15 (see [IBM01]).

Factoring integers can be lucrative. For example, as of September 2002, if you factor the following 174-digit integer then the RSA security company will award you ten thousand dollars! (See [RSA].)

$$\begin{array}{l}18819881292060796383869723946165043980716356337941738270 07\\ 63356422988859715234665485319060606504743045317388011303 39\\ 67161996923212057340318795506569962213051687593076502570 59\end{array}$$

This number is known as RSA-576 since it has 576 digits when written in binary (see Section 3.5.2 for more on binary numbers). RSA constructed this difficult-to-factor number by multiplying together two large primes.

The previous RSA challenge was the 155-digit number

$$\begin{array}{l}10941738641570527421809707322040357612003732945449205990 91\\ 38421314763499842889347847179972578912673324976257528997 81\\ 8337970765372440271467435315933543338 97.\end{array}$$

It was factored on 22 August 1999 by a group of sixteen researchers in four months on a cluster of 292 computers (see [ACD+99]). They found that RSA-155 is the product of the following two 78-digit primes:

$$p = 102639592829741105772054196573991675900716567808038066803341933521790711307779$$
$$q = 106603488380168454820927220360012878679207958575989291522270608237193062808643.$$

3.1.4 The Fundamental Theorem of Arithmetic

We are ready to prove Theorem 3.1.5 using the following idea. Suppose we have two factorizations of n. Using Theorem 3.1.11 we cancel common primes from each factorization, one prime at a time. At the end, we discover that the factorizations must consist of exactly the same primes. The technical details are given below.

Proof. By Proposition 3.1.12, there exist primes $p_1, \ldots, p_d$ such that

$$n = p_1 \cdot p_2 \cdots p_d.$$

Suppose that

$$n = q_1 \cdot q_2 \cdots q_m$$

is another expression of n as a product of primes. Since

$$p_1 \mid n = q_1 \cdot (q_2 \cdots q_m),$$

Euclid's theorem implies that $p_1 = q_1$ or $p_1 \mid q_2 \cdots q_m$. By induction, we see that $p_1 = q_i$ for some i.

Now cancel p_1 and q_i, and repeat the above argument. Eventually, we find that, up to order, the two factorizations are the same. $\square$

3.2 The Sequence of Prime Numbers

This section is concerned with three questions. Are there infinitely many primes? Are there infinitely many primes of the form $ax + b$ for varying $x \in \mathbb{N}$ and fixed integers $a > 1$ and $b \in \mathbb{Z}$? How many primes are there? We first show that there are infinitely many primes, then state Dirichlet's theorem that if $\gcd(a, b) = 1$, then $ax + b$ is a prime for infinitely many values of x. Finally, we discuss the prime number theorem which asserts that there are asymptotically $x/\log(x)$ primes less than x (and we make a connection between this asymptotic formula and the Riemann Hypothesis).

3.2.1 There Are Infinitely Many Primes

Note that each number on the left is prime. Does this continue indefinitely?

$$\begin{aligned}
3 &= 2 + 1\\
7 &= 2 \cdot 3 + 1\\
31 &= 2 \cdot 3 \cdot 5 + 1\\
211 &= 2 \cdot 3 \cdot 5 \cdot 7 + 1\\
2311 &= 2 \cdot 3 \cdot 5 \cdot 7 \cdot 11 + 1
\end{aligned}$$

Theorem 3.2.1 (Euclid). *There are infinitely many primes.*

Proof. Suppose that $p_1, p_2, \ldots, p_n$ are all the primes. If we let

$$N = p_1 p_2 p_3 \cdots p_n + 1,$$

then by Proposition 3.1.12

$$N = q_1 q_2 \cdots q_m$$

with each q_i prime and $m \geq 1$. If $q_1 = p_i$ for some i, then $p_i \mid N$ and $p_i \mid N + 1$, so $p_i \mid 1 = (N + 1) - N$, a contradiction. Thus the prime q_1 is not in the list $p_1, \ldots, p_n$, which is a contradiction. $\square$

For example,

$$2 \cdot 3 \cdot 5 \cdot 7 \cdot 11 \cdot 13 + 1 = 30031 = 59 \cdot 509.$$

Multiplying together the first 6 primes and adding 1 doesn't produce a prime, but it produces an integer that is merely divisible by a new prime.

Joke 3.2.2 (Hendrik Lenstra). *There are infinitely many composite numbers. Proof.* Multiply together the first $n+1$ primes and don't add 1.

3.2.2 The Largest Known Prime

Though Theorem 3.2.1 implies that there are infinitely many primes, it still makes sense to ask the social question "What is the largest *known* prime?"

According to [Cal] the largest known prime, as of September 2002, is the Mersenne prime

$$p = 2^{13466917} - 1,$$

which was discovered in November 2001. (A *Mersenne prime* is a prime of the form $2^q - 1$.) This number has 4053946 decimal digits, so writing it out would fill several large paperback novels.

Euclid's theorem implies that there definitely is a bigger prime; however, nobody has yet found one, proven that they are right, and released their result to the world. Deciding whether or not a number is prime is surprisingly interesting, both as a motivating problem and for applications to cryptography, as we will see in Section 3.5.3 and Chapter 4.

3.2.3 Primes of the Form $ax + b$

Next we turn to primes of the form $ax+b$, where a and b are fixed integers with $a > 1$ and x varies over $\mathbb{N}$. We assume that $\gcd(a, b) = 1$, because otherwise there is no hope that $ax+b$ is prime infinitely often. For example, $2x + 2$ is never prime for $x \in \mathbb{N}$.

Proposition 3.2.3. *There are infinitely many primes of the form* $4x - 1$.

Why might this be true? We list numbers of the form $4x-1$ and underline those that are prime:

$$\underline{3}, \underline{7}, \underline{11}, 15, \underline{19}, \underline{23}, 27, \underline{31}, 35, 39, \underline{43}, \underline{47}, \ldots$$

It is plausible that underlined numbers would continue to appear.

Proof. Suppose $p_1, p_2, \ldots, p_n$ are primes of the form $4x - 1$. Consider the number

$$N = 4p_1p_2 \cdots p_n - 1.$$

Then $p_i \nmid N$ for any i. Moreover, not every prime $p \mid N$ is of the form $4x + 1$; if they all were, then N would be of the form $4x + 1$. Thus there is a $p \mid N$ that is of the form $4x-1$. Since $p \neq p_i$ for any i, we have found a new prime of the form $4x-1$. We can repeat this process indefinitely, so the set of primes of the form $4x-1$ cannot be finite. $\square$

Note that this proof does not work if $4x-1$ is replaced by $4x+1$, since a product of primes of the form $4x-1$ can be of the form $4x+1$. We will give a completely different proof that there are infinitely many primes of the form $4x+1$ using primitive roots in Chapter 5.[2] 2

Example 3.2.4. Set $p_1 = 3$, $p_2 = 7$. Then

$$N = 4 \cdot 3 \cdot 7 - 1 = \underline{83}$$

is a prime of the form $4x-1$. Next

$$N = 4 \cdot 3 \cdot 7 \cdot 83 - 1 = \underline{6971},$$

which is a again a prime of the form $4x-1$. Again:

$$N = 4 \cdot 3 \cdot 7 \cdot 83 \cdot 6971 - 1 = 48601811 = 61 \cdot \underline{796751}.$$

This time 61 is a prime, but it is of the form $4x+1 = 4 \cdot 15 + 1$. However, 796751 is prime and $796751 = 4 \cdot 199188 - 1$. We are unstoppable:

$$N = 4 \cdot 3 \cdot 7 \cdot 83 \cdot 6971 \cdot 796751 - 1 = \underline{5591} \cdot 6926049421.$$

This time the small prime, 5591, is of the form $4x-1$ and the large one is of the form $4x+1$.

Theorem 3.2.5 (Dirichlet). *Let a and b be integers with $\gcd(a,b) = 1$. Then there are infinitely many primes of the form $ax+b$.*

Proofs of this theorem, which use tools from algebraic and analytic number theory, our beyond the scope of this book (for a proof, see [FT93, VIII.4]). [3] 3

---

[2] Noam suggests: Prove there are infinitely many primes of the form $4x+1$ by considering $a = 4(p_1 \cdots p_n)^2 + 1$, which is not divisible by any of $p_1, \ldots, p_n$. Note that modulo any divisor p of a there is a square root of -1, so $p \equiv 1 \pmod 4$. Add this to Chapter 5!

[3] I found this in Apostol's review of Hardy-Wright, 5th edition: *In the Notes to Chapter II the existence of elementary proofs of Dirichlet's theorem on primes in arithmetical progressions should be mentioned and references should be provided. For primes of the form $an+1$, elementary proofs not requiring Dirichlet characters were given by W. Sierpiński [Elementary theory of numbers, PWN, Warsaw, 1964; MR 31 #116] and by I. Niven and H. S. Zuckerman [An introduction to the theory of numbers, fourth edition, Wiley, New York, 1980; MR 81g:10001]. Neither of these books is mentioned in the bibliography*

TABLE 3.1. Values of $\pi(x)$

x	100	200	300	400	500	600	700	800	900	1000
$\pi(x)$	25	46	62	78	95	109	125	139	154	168

3.2.4 How Many Primes are There?

We saw in Section 3.2.1 that there are infinitely many primes. In order to get a sense for just how many primes there are, we consider a few warm-up questions. Then we consider some numerical evidence and state the prime number theorem, which gives an asymptotic answer to our question, and connect this theorem with a form of the Riemann Hypothesis. Our discussion of counting primes in this section is very cursory; for more details read Crandall and Pomerance's excellent book [CP01, §1.1.5].

How[4] many natural numbers are even? Answer: Half of them (but note 4
that the cardinality of the even integers is the same as the cardinality of all integers, because there is a bijection between them). How many natural numbers are of the form $4x-1$? Answer: One fourth of them. How many natural numbers are perfect squares? Answer: Zero percent of all natural numbers, in the sense that the limit of the proportion of perfect squares to all natural numbers converges to 0; more precisely,

$$\lim_{x\to\infty} \frac{\#\{n \in \mathbb{N} : n \leq x \text{ and } n \text{ is a perfect square}\}}{x} = 0,$$

since the numerator is roughly $\sqrt{x}$ and $\lim_{x\to\infty} \frac{\sqrt{x}}{x} = 0$.

Likewise, and we won't prove this here, zero percent of all natural numbers are prime (this follows from Theorem 3.2.7 below). We are thus led to ask the following more precise question: How many positive integers $\leq x$ are perfect squares? Answer: roughly $\sqrt{x}$. In the context of primes, we ask,

Question 3.2.6. How many natural numbers $\leq x$ are prime?

Let

$$\pi(x) = \#\{p \in \mathbb{N} : p \leq x \text{ is a prime}\}.$$

For example,

$$\pi(6) = \#\{2, 3, 5\} = 3.$$

Some values of $\pi(x)$ are given in Table 3.1, and Figure 3.1 contains a graph of $\pi(x)$ for $x < 1000$, which almost looks like a straight line.

Gauss had a serious prime-computing habit; eventually he computed $\pi(3000000)$, though the author doesn't know whether or not Gauss got the right answer, which is 216816. Gauss conjectured the following asymptotic

[4] Note that cardinalities are the same... (Galileo noticed this.)

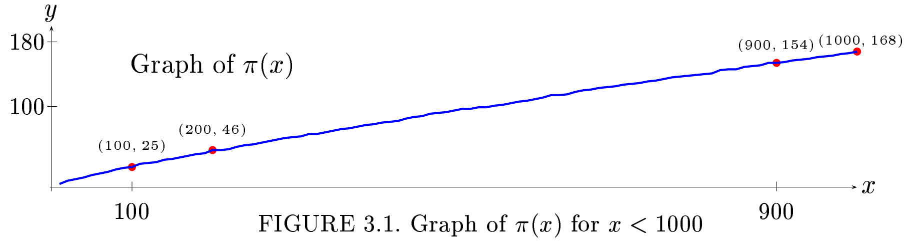


FIGURE 3.1. Graph of $\pi(x)$ for $x < 1000$

TABLE 3.2. Comparison of $\pi(x)$ and $x/(\log(x)-1)$

x	$\pi(x)$	$x/(\log(x)-1)$ (approx)
1000	168	169.2690290604408165186256278
2000	303	302.9888734545463878029800994
3000	430	428.1819317975237043747385740
4000	550	548.3922097278253264133400985
5000	669	665.1418784486502172369455815
6000	783	779.2698885854778626863677374
7000	900	891.3035657223339974352567759
8000	1007	1001.602962794770080754784281
9000	1117	1110.428422963188172310675011
10000	1229	1217.976301461550279200775705

formula for $\pi(x)$, which was later proved independently by Hadamard and Vallée Poussin in 1896 (but won't be proved in this book[5]): 5

Theorem 3.2.7 (Prime Number Theorem). *$\pi(x)$ is asymptotic to $x/\log(x)$, in the sense that*

$$\lim_{x\to\infty} \frac{\pi(x)}{x/\log(x)} = 1.$$

We do nothing more here than motivate this theorem by some some numerical observations.

The theorem implies that $\lim_{x\to\infty} \pi(x)/x = \lim_{x\to\infty} 1/\log(x) = 0$, so for any a,

$$\lim_{x\to\infty} \frac{\pi(x)}{x/(\log(x)-a)} = \lim_{x\to\infty} \frac{\pi(x)}{x/\log(x)} - \frac{a\pi(x)}{x} = 1.$$

Thus $x/(\log(x)-a)$ is also asymptotic to $\pi(x)$ for any a. See [CP01, §1.1.5] for a discussion of why $a = 1$ is the best choice. Table 3.2 compares $\pi(x)$ and $x/(\log(x)-1)$ for several $x < 10000$.

The current world record for counting primes appears to be

$$\pi(4 \cdot 10^{22}) = 783964159847056303858.$$

[5] Give reference.

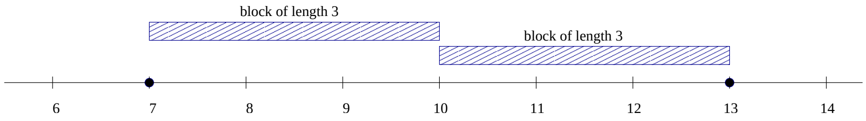


FIGURE 3.2. Visualizing the Mod 3 Congruence Between 7 and 13

The computation of $\pi(4\cdot 10^{22})$ took about 250 days on a 350 Mhz Pentium II; see [GS02] for more details.

The famous Riemann Hypothesis about the location of zeros of the Riemann zeta function $\sum n^{-s}$ is equivalent to the conjecture that

$$\mathrm{Li}(x) = \int_2^x \frac{1}{\log(t)}dt.$$

is an excellent approximation to $\pi(x)$, in the following precise sense (see [CP01, §1.4.1]):

Conjecture 3.2.8 (Equivalent to the Riemann Hypothesis). *For all $x \geq 2.01$,*

$$|\pi(x) - \mathrm{Li}(x)| \leq \sqrt{x}\log(x).$$

Again, we will do nothing more to motivate this here than to give some numerical examples.

Example 3.2.9. Let $x = 4\cdot 10^{22}$. Then

$$\begin{aligned}
\pi(x) &= 783964159847056303858,\\
\mathrm{Li}(x) &= 783964159852157952242.7155276025801473\ldots,\\
|\pi(x)-\mathrm{Li}(x)| &= 5101648384.71552760258014\ldots,\\
\sqrt{x}\log(x) &= 10408633281397.77913344605\ldots,\\
x/(\log(x)-1) &= 783650443647303761503.5237113087392967\ldots.
\end{aligned}$$

3.3 Congruences Modulo n

In this section we define the ring $\mathbb{Z}/n$ of integers modulo n, introduce the Euler φ-function, and relate it to the multiplicative order of certain elements of $\mathbb{Z}/n$.

Definition 3.3.1 (Congruence). Let $a, b \in \mathbb{Z}$ and $n \in \mathbb{N}$. Then a is congruent to b modulo n if $n \mid a-b$. We write $a \equiv b \pmod{n}$.

In other words, a is congruent to b modulo n if we can get from a to b by adding or subtracting copies of n. For example, $13 \equiv 7 \pmod 3$, since $7 = 13 - 3 - 3$, as illustrated in Figure 3.2.

Congruence modulo n is an equivalence relation on $\mathbb{Z}$.

Definition 3.3.2. The *ring of integers modulo* n is the set $\mathbb{Z}/n$ of equivalences classes of integers equipped with its natural ring structure.

Example 3.3.3.

$$\mathbb{Z}/3 = \{\{\ldots,-3,0,3,\ldots\},\{\ldots,-2,1,4,\ldots\},\{\ldots,-1,2,5,\ldots\}\}$$

We use the notation $\mathbb{Z}/n$ because $\mathbb{Z}/n$ is the quotient of the ring $\mathbb{Z}$ by the ideal $n\mathbb{Z}$ of multiples of n. Because $\mathbb{Z}/n$ is the quotient of a ring by an ideal, the ring structure on $\mathbb{Z}$ induces a ring structure on $\mathbb{Z}/n$. We often let a denote the equivalence class of a, when this won't cause confusion. If p is a prime $\mathbb{Z}/p$ is a field (see Exercise 16), which we sometimes also denote by $\mathbb{F}_p$.

It is very easy to derive criteria for divisibility of an integer by 3, 5, 9, and 11 by working modulo n (see Exercise 12). For example,

Proposition 3.3.4. *A number $n \in \mathbb{Z}$ is divisible by* 3 *if and only if the sum of the digits of n is divisible by* 3*.*

Proof. Write

$$n = a + 10b + 100c + \cdots,$$

so the digits of n are a, b, c, etc. Since $10 \equiv 1 \pmod 3$,

$$n = a + 10b + 100c + \cdots \equiv a + b + c + \cdots \pmod 3,$$

from which the proposition follows. □

Definition 3.3.5 (GCD in $\mathbb{Z}/n$). For elements a and b of $\mathbb{Z}/n$, let

$$\gcd(a,b) = \gcd(\tilde{a}, \gcd(\tilde{b}, n)),$$

where $\tilde{a}, \tilde{b} \in \mathbb{Z}$ reduce to a, b, respectively.

It is necessary to check that this is well defined (see Exercise 6).

In order to start solving interesting equations in $\mathbb{Z}/n$, note that it is often possible to cancel a quantity from boths sides of an equation, though sometimes it is not (see Proposition 3.3.11).

Proposition 3.3.6. *If* $\gcd(c,n) = 1$ *and*

$$ac \equiv bc \pmod n,$$

then $a \equiv b \pmod n$*.*

Proof. By definition

$$n \mid ac - bc = (a-b)c.$$

Since $\gcd(n,c) = 1$, it follows that $n \mid a - b$, so

$$a \equiv b \pmod n,$$

as claimed. □

3.3.1 Linear Equations Modulo n

In this section, we are concerned with how to decide whether or not a linear equation of the form $ax \equiv b \pmod{n}$ has a solution modulo n. For example, when a has a multiplicative inverse in $\mathbb{Z}/n$ then $ax \equiv b \pmod{n}$ has a unique solution. Thus it is of interest to determine the units in $\mathbb{Z}/n$, i.e., the elements which have a multiplicative inverse. Finding solutions to $ax \equiv b \pmod{n}$ is the topic of Section 3.5.

We will use complete sets of residues to prove that the units in $\mathbb{Z}/n$ are exactly the $a \in \mathbb{Z}/n$ such that $\gcd(a, n) = 1$.

Definition 3.3.7 (Complete Set of Residues). A subset $R \subset \mathbb{Z}$ of size n whose reductions modulo n are distinct is called a *complete set of residues* modulo n. In other words, a complete set of residues is a choice of representive for each equivalence class in $\mathbb{Z}/n$.

For example,

$$R = \{0, 1, 2, \ldots, n-1\}$$

is a complete set of residues modulo n. When $n = 5$, $R = \{0, 1, -1, 2, -2\}$ is a complete set of residues.

Lemma 3.3.8. *If R is a complete set of residues modulo n and $a \in \mathbb{Z}$ with $\gcd(a, n) = 1$, then $aR = \{ax : x \in R\}$ is also a complete set of residues modulo n.*

Proof. If $ax \equiv ax' \pmod{n}$ with $x, x' \in R$, then Proposition 3.3.6 implies that $x \equiv x' \pmod{n}$. Because R is a complete set of residues, this implies that $x = x'$. Thus the elements of aR have distinct reductions modulo n. It follows, since $\#aR = n$, that aR is a complete set of residues modulo n. $\square$

Proposition 3.3.9. *If $\gcd(a, n) = 1$, then the equation $ax \equiv b \pmod{n}$ has a solution, and the solution is unique modulo n.*

Proof. Let R be a complete set of residues modulo n, so there is a unique element of R that is congruent to b modulo n. By Lemma 3.3.8, aR is also a complete set of residues modulo n, so there is a unique element $ax \in aR$ that is congruent to b modulo n, and we have $ax \equiv b \pmod{n}$. $\square$

Algebraically, this proposition asserts that if $\gcd(a, n) = 1$, then the map $\mathbb{Z}/n \to \mathbb{Z}/n$ given by left multiplication by a is bijective.

Example 3.3.10. Consider $2x \equiv 3 \pmod{7}$. Letting $R = \{0, 1, 2, 3, 4, 5, 6\}$, we have

$$2R = \{0, 2, 4, 6, 8 \equiv 1, 10 \equiv 3, 12 \equiv 5\},$$

so $2 \cdot 5 \equiv 3 \pmod{7}$.

When $\gcd(a, n) \neq 1$, then the equation $ax \equiv b \pmod{n}$ may or may not have a solution. For example, $2x \equiv 1 \pmod{4}$ has no solution, but $2x \equiv 2 \pmod{4}$ does, and in fact it has more than one ($x = 1$ and $x =$

3). Generalizing Proposition 3.3.9 we obtain the following more general criterion for solvability.

Proposition 3.3.11. *The equation* $ax \equiv b \pmod{n}$ *has a solution if and only if* $\gcd(a,n)$ *divides* b.

Proof. Let $g = \gcd(a,n)$. If there is a solution x to the equation, then $n \mid (ax - b)$. Since $g \mid n$ and $g \mid a$, it follows that $g \mid b$.

Conversely, suppose that $g \mid b$. Then $n \mid (ax - b)$ if and only if

$$\frac{n}{g} \mid \left(\frac{a}{g}x - \frac{b}{g}\right).$$

Thus $ax \equiv b \pmod{n}$ has a solution if and only if $\frac{a}{g}x \equiv \frac{b}{g} \pmod{\frac{n}{g}}$ has a solution. By Proposition 3.3.9, this latter equation does have a solution. □

3.3.2 Fermat's Little Theorem

Definition 3.3.12 (Order of an Element). Let $n \in \mathbb{N}$ and $x \in \mathbb{Z}$ with $\gcd(x,n) = 1$. The *order* of x modulo n is the smallest $m \in \mathbb{N}$ such that

$$x^m \equiv 1 \pmod{n}.$$

To show that the definition makes sense, we verify that such an m exists. Consider $x, x^2, x^3, \ldots$ modulo n. There are only finitely many residue classes modulo n, so we must eventually find two integers i, j with $i < j$ such that

$$x^j \equiv x^i \pmod{n}.$$

Since $\gcd(x,n) = 1$, Proposition 3.3.6 implies that we can cancel x's and conclude that

$$x^{j-i} \equiv 1 \pmod{n}.$$

Definition 3.3.13 (Euler Phi function). For $n \in \mathbb{N}$, let

$$\varphi(n) = \#\{a \in \mathbb{N} : a \leq n \text{ and } \gcd(a,n) = 1\}.$$

For example,

$$\begin{aligned}
\varphi(1) &= \#\{1\} = 1,\\
\varphi(2) &= \#\{1\} = 1,\\
\varphi(5) &= \#\{1,2,3,4\} = 4,\\
\varphi(12) &= \#\{1,5,7,11\} = 4.
\end{aligned}$$

Also, if p is any prime number then

$$\varphi(p) = \#\{1, 2, \ldots, p-1\} = p - 1.$$

In Section 3.4.1, we will prove that φ is a multiplicative function. This will yield an easy way to compute $\varphi(n)$ in terms of the prime factorization of n.

Theorem 3.3.14 (Fermat's Little Theorem). *If* $\gcd(x,n) = 1$, *then*

$$x^{\varphi(n)} \equiv 1 \pmod{n}.$$

Proof. Let

$$P = \{a : 1 \le a \le n \text{ and } \gcd(a,n) = 1\}.$$

In the same way that we proved Lemma 3.3.8, we see that the reductions modulo n of the elements of xP are the same as the reductions of the elements of P. Thus

$$\prod_{a\in P}(xa) \equiv \prod_{a\in P} a \pmod{n},$$

since the products are over the same numbers modulo n. Now cancel the a's on both sides to get

$$x^{\#P} \equiv 1 \pmod{n},$$

as claimed. □

Note that $\varphi(n)$ is not, of course, necessarily equal to the order of x modulo n. For example, if $x = 1$ and $n > 2$, then x has order 1, but $\varphi(n) > 1$. Theorem 3.3.14 only implies that $\varphi(n)$ is a multiple of the order of x.

Fermat's Little Theorem has the following group theoretic interpretation. The set of units in $\mathbb{Z}/n$ is a group

$$(\mathbb{Z}/n)^{\times} = \{a \in \mathbb{Z}/n : \gcd(a,n) = 1\}.$$

which has order $\varphi(n)$. Theorem 3.3.14 asserts that the order of an element of $(\mathbb{Z}/n)^{\times}$ divides the order $\varphi(n)$ of $(\mathbb{Z}/n)^{\times}$. This is a special case of the more general fact that if G is a finite group and $g \in G$, then the order of g divides the cardinality of G.

3.3.3 Wilson's Theorem

The following result, from the 1770s, is called "Wilson's Theorem" (though it was first proved by Lagrange).

Proposition 3.3.15 (Wilson's Theorem). *An integer* $p > 1$ *is prime if and only if* $(p-1)! \equiv -1 \pmod{p}$.

For example, if $p = 3$, then $(p-1)! = 2 \equiv -1 \pmod{3}$. If $p = 17$, then

$$(p-1)! = 20922789888000 \equiv -1 \pmod{17}.$$

But if $p = 15$, then

$$(p-1)! = 87178291200 \equiv 0 \pmod{15},$$

so 15 is composite. Thus Wilson's theorem could be viewed as a primality test, though, from a computational point of view, it is probably the *least efficient* primality test since computing $(n-1)!$ takes far more steps than checking for prime divisors of n up to $\sqrt{n}$.

Proof. We first assume that p is prime and prove that $(p-1)! \equiv -1 \pmod{p}$. If $a \in \{1, 2, \ldots, p-1\}$ then the equation

$$ax \equiv 1 \pmod{p}$$

has a unique solution $a' \in \{1, 2, \ldots, p-1\}$. If $a = a'$, then $a^2 \equiv 1 \pmod{p}$, so $p \mid a^2 - 1 = (a-1)(a+1)$, so $p \mid (a-1)$ or $p \mid (a+1)$, so $a \in \{1, -1\}$. We can thus pair off the elements of $\{2, 3, \ldots, p-2\}$, each with their inverse. Thus

$$2 \cdot 3 \cdot \cdots \cdot (p-2) \equiv 1 \pmod{p}.$$

Multiplying both sides by $p-1$ proves that $(p-1)! \equiv -1 \pmod{p}$.

Next we assume that $(p-1)! \equiv -1 \pmod{p}$ and prove that p must be prime. Suppose not, so that $p \geq 4$ is a composite number. Let ℓ be a prime divisor of p. Then $\ell < p$, so $\ell \mid (p-1)!$. Also, by assumption,

$$\ell \mid p \mid ((p-1)! + 1).$$

This is a contradiction, because a prime can't divide a number a and also divide $a+1$, since it would then have to divide $(a+1) - a = 1$. $\square$

Example 3.3.16. We illustrate the key step in the above proof in the case $p = 17$. We have

$$2 \cdot 3 \cdots 15 = (2 \cdot 9) \cdot (3 \cdot 6) \cdot (4 \cdot 13) \cdot (5 \cdot 7) \cdot (8 \cdot 15) \cdot (10 \cdot 12) \cdot (14 \cdot 11) \equiv 1 \pmod{17},$$

where we have paired up the numbers a, b for which $ab \equiv 1 \pmod{17}$.

3.4 The Chinese Remainder Theorem

In this section we prove the Chinese Remainder Theorem, which gives conditions under which a system of linear equations is guaranteed to have a solution.

A 4th century Chinese mathematician asked:

Question 3.4.1. There is a quantity whose number is unknown. Repeatedly divided by 3, the remainder is 2; by 5 the remainder is 3; and by 7 the remainder is 2. What is the quantity?

In modern notation, Question 3.4.1 asks us to find a positive integer solution to the following system of three equations:

$$\begin{aligned} x &\equiv 2 \pmod{3} \\ x &\equiv 3 \pmod{5} \\ x &\equiv 2 \pmod{7} \end{aligned}$$

The Chinese Remainder Theorem asserts that a solution exists, and the proof gives a method to find one. (See Section 3.5 for the necessary algorithms.)

Theorem 3.4.2 (Chinese Remainder Theorem). *Let $a, b \in \mathbb{Z}$ and $n, m \in \mathbb{N}$ such that $\gcd(n, m) = 1$. Then there exists $x \in \mathbb{Z}$ such that*

$$\begin{aligned} x &\equiv a \pmod{m}, \\ x &\equiv b \pmod{n}. \end{aligned}$$

Moreover x is unique modulo mn.

Proof. If we can solve for t in the equation

$$a + tm \equiv b \pmod{n},$$

then $x = a + tm$ will satisfy both congruences. To see that we can solve, subtract a from both sides and use Proposition 3.3.9 together with our assumption that $\gcd(n, m) = 1$ to see that there is a solution.

For uniqueness, suppose that x and y solve both congruences. Then $z = x - y$ satisfies $z \equiv 0 \pmod{m}$ and $z \equiv 0 \pmod{n}$, so $m \mid z$ and $n \mid z$. Since $\gcd(n, m) = 1$, it follows that $nm \mid z$, so $x \equiv y \pmod{nm}$. □

Now we can answer Question 3.4.1. First, we use Theorem 3.4.2 to find a solution to the pair of equations

$$\begin{aligned} x &\equiv 2 \pmod{3} \\ x &\equiv 3 \pmod{5}. \end{aligned}$$

Set $a = 2$, $b = 3$, $m = 3$, $n = 5$. Step 1 is to find a solution to $t \cdot 3 \equiv 3 - 2 \pmod{5}$. A solution is $t = 2$. Then $x = a + tm = 2 + 2 \cdot 3 = 8$. Since any x' with $x' \equiv x \pmod{15}$ is also a solution to those two equations, we can solve all three equations by finding a solution to the pair of equations

$$\begin{aligned} x &\equiv 8 \pmod{15} \\ x &\equiv 2 \pmod{7}. \end{aligned}$$

Again, we find a solution to $t \cdot 15 \equiv 2 - 8 \pmod{7}$. A solution is $t = 1$, so

$$x = a + tm = 8 + 15 = 23.$$

Note that there are other solutions. Any $x' \equiv x \pmod{3 \cdot 5 \cdot 7}$ is also a solution; e.g., $23 + 3 \cdot 5 \cdot 7 = 128$.

3.4.1 Multiplicative Functions

Definition 3.4.3. A function $f : \mathbb{N} \to \mathbb{Z}$ is *multiplicative* if, whenever $m, n \in \mathbb{N}$ and $\gcd(m, n) = 1$, we have

$$f(mn) = f(m) \cdot f(n).$$

Recall that the *Euler φ-function* is

$$\varphi(n) = \#\{a : 1 \leq a \leq n \text{ and } \gcd(a, n) = 1\}.$$

Proposition 3.4.4. *φ is a multiplicative function.*

Proof. Suppose that $m, n \in \mathbb{N}$ and $\gcd(m, n) = 1$. Consider the map

$$r : (\mathbb{Z}/mn)^\times \to (\mathbb{Z}/m)^\times \times (\mathbb{Z}/n)^\times.$$

defined by

$$r(c) = (c \bmod m,\ c \bmod n).$$

We first show that r is injective. If $r(c) = r(c')$, then $m \mid c - c'$ and $n \mid c - c'$, so, since $\gcd(n, m) = 1$, $nm \mid c - c'$, so $c = c'$ as elements of $(\mathbb{Z}/mn)^\times$.

Next we show that r is surjective. Given a and b with $\gcd(a, m) = 1$ and $\gcd(b, n) = 1$, Theorem 3.4.2 implies that there exists c with $c \equiv a \pmod{m}$ and $c \equiv b \pmod{n}$. We may assume that $1 \leq c \leq nm$, and since $\gcd(a, m) = 1$ and $\gcd(b, n) = 1$, we must have $\gcd(c, nm) = 1$. Thus $r(c) = (a, b)$.

Because r is a bijection, the set on the left has the same size as the product set on the right. Thus

$$\varphi(mn) = \varphi(m) \cdot \varphi(n).$$

□

For an alternative proof see Exercise 22

The proposition makes it easier to compute $\varphi(n)$. For example,

$$\varphi(12) = \varphi(2^2) \cdot \varphi(3) = 2 \cdot 2 = 4.$$

Also, for $n \geq 1$, we have

$$\varphi(p^n) = p^n - \frac{p^n}{p} = p^n - p^{n-1} = p^{n-1}(p-1),$$

since $\varphi(p^n)$ is the number of numbers less than p^n minus the number of those that are divisible by p. Thus, e.g.,

$$\varphi(389 \cdot 11^2) = 388 \cdot (11^2 - 11) = 388 \cdot 110 = 42680.$$

For a discussion of a relation between computing $\varphi(n)$ and factoring n in certain cases, see Section 4.3.1.

3.5 Quickly Computing Inverses and Huge Powers

This section is about how to solve $ax \equiv 1 \pmod{n}$ when we know it has a solution, and how to efficiently compute $a^m \pmod{n}$. We also discuss a simple probabilistic primality test that relies on being able to compute $a^m \pmod{n}$ quickly. All three of these algorithms are of fundamental importance in Chapter 4, since they lie at the heart of the Diffie-Hellman and RSA public-key cryptosystems.

3.5.1 How to Solve $ax \equiv 1 \pmod{n}$

Suppose $a, n \in \mathbb{N}$ with $\gcd(a, n) = 1$. Then by Proposition 3.3.9 the equation $ax \equiv 1 \pmod{n}$ has a unique solution. How can we find it?

Proposition 3.5.1. *Suppose $a, b \in \mathbb{Z}$ and $\gcd(a, b) = d$. Then there exists $x, y \in \mathbb{Z}$ such that*

$$ax + by = d.$$

Remark 3.5.2. If $e = cd$ is a multiple of d, then $cax + cby = cd = e$, so e can also be written in terms of a and b.

We won't formally prove Proposition 3.5.1, but instead we show how to find x and y in practice. To use this proposition to solve $ax \equiv 1 \pmod{n}$, use that $\gcd(a, n) = 1$ to find x and y such that $ax + ny = 1$. Then

$$ax \equiv 1 \pmod{n}.$$

Suppose $a = 5$ and $b = 7$. The steps of the Euclidean gcd algorithm (Algorithm 3.1.8) are:

$$\begin{aligned}
\underline{7} &= 1 \cdot \underline{5} + \underline{2} && \text{so } \underline{2} = \underline{7} - \underline{5} \\
\underline{5} &= 2 \cdot \underline{2} + \underline{1} && \text{so } \underline{1} = \underline{5} - 2 \cdot \underline{2} = \underline{5} - 2(\underline{7} - \underline{5}) = 3 \cdot \underline{5} - 2 \cdot \underline{7}
\end{aligned}$$

On the right, we have backsubstituted in order to write each partial remainder as a linear combination of a and b. In the last step, we obtain $\gcd(a,b)$ as a linear combination of a and b, as desired.

That example wasn't too complicated, so we try a longer one. Let $a = 130$ and $b = 61$. We have

$$\begin{aligned} \underline{130} &= 2\cdot\underline{61}+\underline{8} & \underline{8} &= \underline{130} - 2\cdot\underline{61} \\ \underline{61} &= 7\cdot\underline{8}+\underline{5} & \underline{5} &= -7\cdot\underline{130} + 15\cdot\underline{61} \\ \underline{8} &= 1\cdot\underline{5}+\underline{3} & \underline{3} &= 8\cdot\underline{130} - 17\cdot\underline{61} \\ \underline{5} &= 1\cdot\underline{3}+\underline{2} & \underline{2} &= -15\cdot\underline{130} + 32\cdot\underline{61} \\ \underline{3} &= 1\cdot\underline{2}+\underline{1} & \underline{1} &= 23\cdot\underline{130} - 49\cdot\underline{61} \end{aligned}$$

Thus $x = 23$ and $y = -49$ is a solution to $130x + 61y = 1$.

For the purpose of solving $ax \equiv 1 \pmod{n}$, it is sufficient to find any solution to $ax + by = d$. In fact, there are always infinitely many solutions to this equation; if x, y is a solution to

$$ax + by = d,$$

then for any $c \in \mathbb{Z}$,

$$a\left(x + c\cdot\frac{b}{d}\right) + b\left(y - c\cdot\frac{a}{d}\right) = d,$$

is also a solution. Moreover, all solutions are of the above form for some c.

Example 3.5.3. Solve $17x \equiv 1 \pmod{61}$. First, we use the Euclidean algorithm to find x, y such that $17x + 61y = 1$:

$$\begin{aligned} \underline{61} &= 3\cdot\underline{17}+\underline{10} & \underline{10} &= \underline{61} - 3\cdot\underline{17} \\ \underline{17} &= 1\cdot\underline{10}+\underline{7} & \underline{7} &= -\underline{61} + 4\cdot\underline{17} \\ \underline{10} &= 1\cdot\underline{7}+\underline{3} & \underline{3} &= 2\cdot\underline{61} - 7\cdot\underline{17} \\ \underline{3} &= 2\cdot\underline{3}+\underline{1} & \underline{1} &= -5\cdot\underline{61} + 18\cdot\underline{17} \end{aligned}$$

Thus $17\cdot 18 + 61\cdot(-5) = 1$ so $x = 18$ is a solution to $17x \equiv 1 \pmod{61}$.

To simplify this process, we view it algebraically as follows. Define a homomorphism $\varphi : \mathbb{Z}\times\mathbb{Z} \to \mathbb{Z}$ by $\varphi(x,y) = xa + yb$. Our goal is to find (x,y) such that $\varphi(x,y) = \gcd(a,b)$. We have $\varphi(1,0) = a$ and $\varphi(0,1) = b$. Each step of the Euclidean algorithm produces a new element of $\mathbb{Z}\times\mathbb{Z}$ that maps to the remainder at that step, and in the end we obtain an (x,y) that maps to $\gcd(a,b)$. We illustrate this with $a = 61$ and $b = 17$:

$$\begin{aligned}
(1,0) &\mapsto 61 & & \\
(0,1) &\mapsto 17 & &\text{multiply by } -3\\
(1,-3) &\mapsto 10 & &-1\\
(-1,4) &\mapsto 7 & &-1\\
(2,-7) &\mapsto 3 & &-2\\
(-5,18) &\mapsto 1. & &
\end{aligned}$$

Thus $61 \cdot (-5) + 17 \cdot 18 = 1$. The parenthesis, commas, and $\mapsto$ symbol are redundant. The following example illustrates writing 1 in terms of 136 and 75 with minimal distracting notation.

x	y	$\varphi(x,y)$	multiple
1	0	136	
0	1	75	-1
1	-1	61	-1
-1	2	14	-4
5	-9	5	-2
-11	20	4	-1
16	-29	1	

Thus $136 \cdot 16 + 75 \cdot (-29) = 1$.

3.5.2 How to Compute $a^m \pmod{n}$

Let a and n be integers, and m a nonnegative integer. In this section we describe an efficient algorithm to compute $a^m \pmod{n}$. For the cryptography applications in Chapter 4, m will have hundreds of digits.

The naive approach to computing $a^m \pmod{n}$ is to simply compute $a^m = a \cdot a \cdots a \pmod{n}$ by repeatedly multiplying by a and reducing modulo m. Note that after each arithmetic operation is completed, we reduce the result modulo n so that the sizes of the numbers involved don't explode. Nonetheless, this algorithm is horribly inefficient because it takes $m - 1$ multiplications, which is out of the question when m has hundreds of digits.

A much more efficient algorithm for computing $a^m \pmod{n}$ involves writing m in binary, then expressing a^m as a product of expressions a^{2^i}, for various i. These latter expressions can be computed by repeatedly squaring a^{2^i}. This more clever algorithm is not "simpler", but it is vastly more efficient since the number of operations needed grows with the number of binary digits of m, whereas with the naive algorithm above the number of operations is $m - 1$.

Algorithm 3.5.4 (Writing a number in binary). Let m be a non-negative integer. This algorithm writes m in binary, so it finds $\varepsilon_i \in \{0,1\}$ such that $m = \sum_{i=0}^{r} \varepsilon_i 2^i$ with each $\varepsilon_i \in \{0,1\}$. If m is odd, then $\varepsilon_0 = 1$, otherwise $\varepsilon_0 = 0$. Replace m by $\lfloor \frac{m}{2} \rfloor$. If the new m is odd then $\varepsilon_1 = 1$, otherwise $\varepsilon_1 = 0$. Keep repeating until $m = 0$.

Algorithm 3.5.5 (Compute $a^m \pmod{n}$). Let a and n be integers and m a nonnegative integer. This algorithm computes a^m modulo n. Write m in binary using Algorithm 3.5.4. Then

$$a^m = \prod_{\varepsilon_i = 1} a^{2^i} \pmod{n}.$$

To compute a^m compute a, a^2, $a^{2^2} = (a^2)^2$, $a^{2^3} = (a^{2^2})^2$, etc., up to a^{2^r}, where $r+1$ is the number of binary digits of m. Then multiply together the a^{2^i} such that $\varepsilon_i = 1$, always working modulo n.

For example, we can compute the last 2 digits of 6^{91}, by finding $6^{91} \pmod{100}$. Make a table whose first column, labeled i, contains 0, 1, 2, etc. The second column, labeled m, is got by dividing the entry above it by 2 and taking the integer part of the result. The third column, labeled ε_i, records whether or not the second column is odd. The forth column is computed by squaring, modulo $n = 100$, the entry above it.

i	m	ε_i	6^{2^i} mod 100
0	91	1	6
1	45	1	36
2	22	0	96
3	11	1	16
4	5	1	56
5	2	0	36
6	1	1	96

We have

$$6^{91} \equiv 6^{2^6} \cdot 6^{2^4} \cdot 6^{2^3} \cdot 6^2 \cdot 6 \equiv 96 \cdot 56 \cdot 16 \cdot 36 \cdot 6 \equiv 56 \pmod{100}.$$

That's a lot easier than multiply 6 by itself 91 times.

3.5.3 A Probabilistic Primality Test

Theorem 3.5.6. *An integer $p > 1$ is prime if and only if for* every $a \not\equiv 0 \pmod{p}$,

$$a^{p-1} \equiv 1 \pmod{p}.$$

Proof. If p is prime, then the statement follows from Proposition 3.3.15. If p is composite, then there is a divisor a of p with $a \neq 1, p$. If $a^{p-1} \equiv$

$1 \pmod{p}$, then $p \mid a^{p-1} - 1$. Since $a \mid p$, $a \mid a^{p-1} - 1$ hence $a \mid 1$, a contradiction. □

Suppose $n \in \mathbb{N}$. Using this theorem and Algorithm 3.5.5, we can either quickly prove that n is not prime, or convince ourselves that n probably is prime. For example, if $2^{n-1} \not\equiv 1 \pmod{n}$, then we have proved that n is not prime. On the other hand, if $a^{p-1} \equiv 1 \pmod{p}$ for a couple of a, it "seems likely" that n is prime.

Example 3.5.7. Is $p = 323$ prime? We compute $2^{322} \pmod{323}$. Making a table as above, we have

i	m	ε_i	$2^{2^i} \bmod 323$
0	322	0	2
1	161	1	4
2	80	0	16
3	40	0	256
4	20	0	290
5	10	0	120
6	5	1	188
7	2	0	137
8	1	1	35

Thus

$$2^{322} \equiv 4 \cdot 188 \cdot 35 \equiv 157 \pmod{323},$$

so 323 is not prime. In fact, $323 = 17 \cdot 19$.

It's possible to prove that a large number is composite, but yet be unable to easily find a factorization! For example if

$$n = 95468093486093450983409583409850934850938459083,$$

then $2^{n-1} \not\equiv 1 \pmod{n}$, so n is composite. We could verify with some work that n is composite with pencil and paper, but factoring n by hand would be extremely difficult.

3.5.4 A Polynomial Time Primality Test

Though the practical method for deciding primality with high probability discussed above is very efficient in practice, it was for a long time an open problem to give an algorithm that decides whether or not any integer is prime in time bounded by a polynomial in the number of digits of the integer. Three Indian mathematicians, Agrawal, Kayal, and Saxena, recently found the first ever polynomial-time primality test. See [AKS02] and also [Ber] for a concise exposition of their clever idea.

Exercises

3.1 Let p be a prime number and r an integer such that $1 \leq r < p$. Prove that p divides the binomial coefficient

$$\frac{p!}{r!(p-r)!}.$$

You may not assume that this coefficient is a integer.

3.2 Compute the following gcd's using a pencil and the Euclidean algorithm:

$$\gcd(15, 35), \quad \gcd(247, 299), \quad \gcd(51, 897), \quad \gcd(136, 304)$$

3.3 (a) Show that 2 is irreducible in the ring $\mathbb{Z}[\sqrt{-5}]$. [Hint: Suppose $2 = (a + b\sqrt{-5})(c + d\sqrt{-5})$, take norms, and apply Theorem 3.1.5.]

(b) Show that $(1 + \sqrt{-5})$ is irreducible in $\mathbb{Z}[\sqrt{-5}]$. [Hint: Suppose $(1 + \sqrt{-5}) = (a + b\sqrt{-5})(c + d\sqrt{-5})$ and take norms.]

3.4 What was the most recent prime year?

3.5 Use the Euclidean algorithm to find integers $x, y \in \mathbb{Z}$ such that

$$2261x + 1275y = 17.$$

3.6 Prove that Definition 3.3.5 is well defined. That is, $\gcd(\tilde{a}, \gcd(\tilde{b}, n))$ doesn't depend on the choice of lifts $\tilde{a}, \tilde{b} \in \mathbb{Z}$.

3.7 Let $f(x) \in \mathbb{Z}[x]$ be a polynomial with integer coefficients. Formulate a conjecture about when the set $\{f(a) : a \in \mathbb{Z} \text{ and } f(a) \text{ is prime}\}$ is infinite. Give computational evidence for your conjecture.

3.8 Is it "easy" or "hard" for a computer to compute the gcd of two random 2000-digit numbers?

3.9 Prove that there are infinitely many primes of the form $6x - 1$.

3.10 (a) Let y be the current year (e.g., 2002). Use a computer to compute

$$\pi(y) = \#\{ \text{ primes } p \leq y\}.$$

(b) The prime number theorem predicts that $\pi(x)$ is asymptotic to $x/\log(x)$. How close is $\pi(y)$ to $y/\log(y)$, where y is as in (a)?

3.11 Find complete sets of residues modulo 7, all of whose elements are (a) nonnegative, (b) odd, (c) even, (d) prime.

3.12 Find rules for divisiblity of an integer by 5, 9, and 11, and prove each of these rules using arithmetic modulo n.

3.13 Find an integer x such that $37x \equiv 1 \pmod{101}$.

3.14 What is the order of 5 modulo 37?

3.15 Let $n = \varphi(20!) = 416084687585280000$. Compute the prime factorization of n using the multiplicative property of φ.

3.16 Let p be a prime. Prove that $\mathbb{Z}/p$ is a field.

3.17 Find an $x \in \mathbb{Z}$ such that $x \equiv -4 \pmod{17}$ and $x \equiv 3 \pmod{23}$.

3.18 Compute the last two digits of 6^{66}.

3.19 Find a number a such that $0 \leq a < 111$ and

$$(102^{70} + 1)^{35} \equiv a \pmod{111}.$$

3.20 Prove that if $n > 4$ is composite then

$$(n-1)! \equiv 0 \pmod{n}.$$

3.21 For what values of n is $\varphi(n)$ odd?

3.22 Prove that φ is multiplicative as follows. Show that the natural map $\mathbb{Z}/mn \to \mathbb{Z}/m \times \mathbb{Z}/n$ is an injective map of rings, hence bijective by counting, then look at unit groups.

3.23 Suppose n is a random 1000 digit number. Do you think computing $\varphi(n)$ is relatively easy or extremely difficult?

3.24 Let $\varphi : \mathbb{N} \to \mathbb{N}$ be the Euler φ function.

(a) Find all natural numbers n such that $\varphi(n) = 1$.

(b) Do there exist natural numbers m and n such that $\varphi(mn) \neq \varphi(m) \cdot \varphi(n)$?

4
Public-Key Cryptography

4.1 The Diffie-Hellman Key Exchange

I recently watched a TV show called *La Femme Nikita* about a skilled women named Nikita, who is forced to be an agent for the anti-terrorist organization Section One. Nikita has strong feelings for fellow agent Michael, and she mosts trust Walter, Section One's gadgets and explosives expert. Often Nikita's worst enemies are her superiors and coworkers at Section One.

The synopsis for a third season episode (which I haven't watched) is as follows:

> PLAYING WITH FIRE
>
> On a mission to secure detonation chips from a terrorist organization's heavily armed base camp, Nikita is captured as a hostage by the enemy. Or so it is made to look. Michael and Nikita have actually created the scenario in order to secretly rendezvous with each other. The ruse works, but when Birkoff [Section One's master hacker] accidentally discovers encrypted messages between Michael and Nikita sent with Walter's help, Birkoff is forced to tell Madeline. Suspecting that Michael and Nikita may be planning a coup d'tat, Operations and Madeline

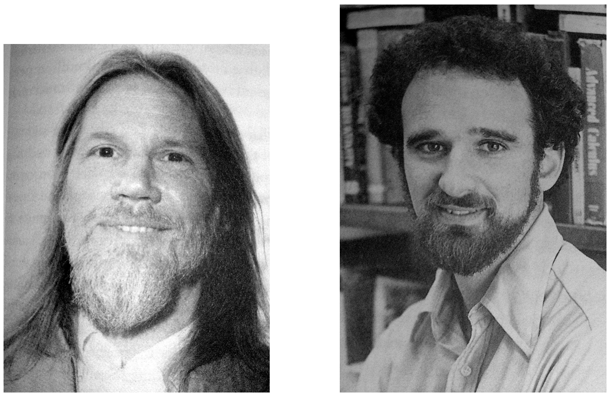


FIGURE 4.1. Diffie and Hellman (photos from [Sin99])

> use a second team of operatives to track Michael and Nikita's next secret rendezvous... killing them if necessary.

What sort of encryption might Walter have helped them to use? I let my imagination run free, and this is what I came up with. After being captured at the base camp, Nikita is given a phone by her captors, in hopes that she'll use it and they'll be able to figure out what she is really up to. Everyone is eagerly listening in on her calls.

Nikita remembers a conversation with Walter about the first public key-exchange protocol, the "Diffie-Hellman key exchange". She remembers that it allows two people to agree on a secret key in the presence of eavesdroppers. Moreover, Walter mentioned that though Diffie-Hellman was the first ever public-key exchange system, it is *still* in common use today (e.g., in `ssh` and `SSL`). It must be good. Perfect!

Nikita pulls out her handheld computer and phone, calls up Michael, and they do the following:

1. Together they choose a big prime number p and a number g with $1 < g < p$.

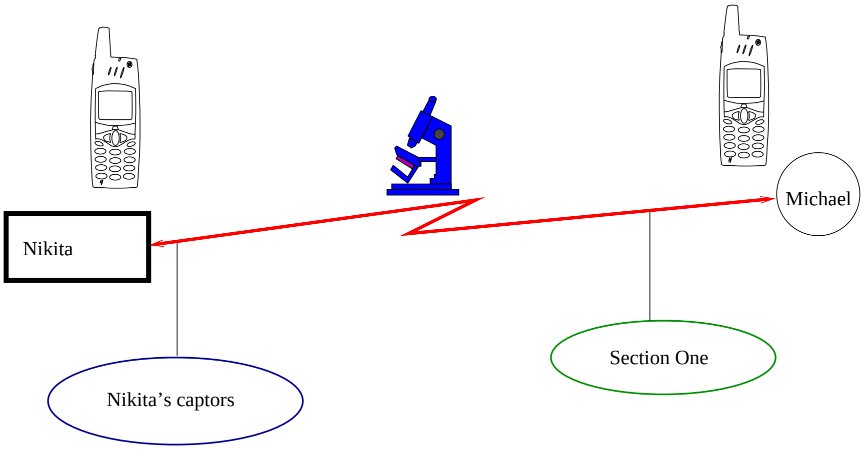
2. Nikita *secretely* chooses an integer n.

3. Michael *secretely* choses an integer m.

4. Nikita tells Michael $ng \pmod p$ (the remainder of ng reduced modulo p).

5. Michael tells $mg \pmod p$ to Nikita.

6. The "secret key" is $s = nmg \pmod p$, which both Nikita and Michael can easily compute.



Here's a very simple example with small numbers that illustrates what Michael and Nikita do. (They really used 200 digit numbers.)

1. $p = 97$, $g = 5$
2. $n = 31$
3. $m = 95$
4. $ng \equiv 58 \pmod{97}$
5. $mg \equiv 87 \pmod{97}$
6. $s = nmg = 78 \pmod{97}$

Nikita and Michael are foiled because everyone easily figures out s:

1. Everyone knows p, g, $ng \pmod{p}$, and $mg \pmod{p}$.
2. Using the very fast Euclidean algorithm, anyone can easily find $a, b \in \mathbb{Z}$ such that $ag + bp = 1$, which exist because $\gcd(g, p) = 1$.
3. Then $ang \equiv n \pmod{p}$, so everyone knows Nikita's secret key n, and hence can find s just as easily as she did.

To taunt her, Nikita's captors give her the Math Review of Diffie and Hellman's 1976 paper "New Directions in Cryptography":

> "The authors discuss some recent results in communications theory [...] The first [method] has the feature that an unauthorized 'eavesdropper' will find it computationally infeasible to decipher the message [...] They propose a couple of techniques for implementing the system, but the reviewer was unconvinced."

Night darkens her cell as Nikita reflects on what has happened. Upon realizing that she misremembered how the system works, she phones Michael and they do the following:

1. Together Michael and Nikita choose a 200-digit (pseudo-)prime p and a number g with $1 < g < p$.

2. Nikita *secretly* chooses an integer n.
3. Michael *secretly* chooses an integer m.
4. Nikita computes $g^n \pmod{p}$ on her handheld computer and tells Michael the resulting number over the phone. (She is surprised that her handheld computer finds $g^n \pmod{p}$ quickly, even though n is very large. How it does this was described in Section 3.5.)
5. Michael tells Nikita $g^m \pmod{p}$.
6. The secret key is then
$$s \equiv (g^n)^m \equiv (g^m)^n \equiv g^{nm} \pmod{p}.$$

Here's a simplified example that illustrates what they did, but which involves only relatively simple arithmetic.

1. $p = 97$, $g = 5$
2. $n = 31$
3. $m = 95$
4. $g^n \equiv 7 \pmod{p}$
5. $g^m \equiv 39 \pmod{p}$
6. $s \equiv (g^n)^m \equiv 14 \pmod{p}$

4.1.1 The Discrete Log Problem

Nikita communicates with Michael by encrypting everything using their agreed upon secret key. In order to understand the conversation, the eavesdroppers need s, but it takes a long time to compute s given only p, g, g^n, and g^m. One way would be to compute n from knowledge of g and g^n; this is possible, but appears to be "computationally infeasible", in the sense that it would take too long to be practical.

Let a, b, and n be real numbers with $a, b > 0$ and $n \geq 0$. Recall that
$$\log_b(a) = n \text{ if and only if } a = b^n.$$
The $\log_b$ function is used in algebra to solve the following problem: Given a base b and a power a of b, find an exponent n such that
$$a = b^n.$$
That is, given $a = b^n$ and b, find n.

Example 4.1.1. The number $a = 19683$ is the nth power of $b = 3$ for some n. With a calculator we quickly find that
$$n = \log_3(19683) = \log(19683)/\log(3) = 9.$$
A calculator computes an approximation for $\log(x)$ quickly by computing a partial sum of a rapidly-converging infinite series.

The discrete log problem is the analogue of this problem but in any finite ("discrete") group:

Problem 4.1.2 (Discrete Log Problem). Let G be a finite group, e.g., $G = (\mathbb{Z}/p)^\times$. Given $b \in G$ and a power a of b, find the smallest positive integer n such that $b^n = a$. Thus the discrete log problem is the problem of computing $n = \log_b(a)$ for $a, b \in G$.

As far as we know, computing discrete logarithms is very time consuming in practice. Over the years, many people have been very motivated to try. For example, if Nikita's captors could efficiently solve Problem 4.1.2, then they could read the messages she exchanges with Michael. Unfortunately, we have no proofs that computing discrete logarithms on a classical computer is difficult. In contrast, Peter Shor [Sho97] showed that quantum computers of significant complexity can solve the discrete logarithm problem in time bounded by a polynomial in the number of digits of $\#G$.

It's easy to give a *slow* algorithm that inefficiently solves the discrete log problem. Simply try b^1, b^2, b^3, etc., until we find an exponent n such that $b^n = a$. For example, suppose $a = 18$, $b = 5$, and $p = 23$. We have

$$b^1 = 5,\ b^2 = 2,\ b^3 = 10,\ \ldots,\ b^{12} = 18,$$

so $n = 12$.

When p is large, computing the discrete log this way soon becomes impractical, because doubling the number of digits of the modulus makes the computation take much longer.

4.1.2 Realistic Example

In this section we present an example that uses bigger numbers.

Let $p = 93450983094850938450983409623$ and $g = -2 \in (\mathbb{Z}/p)^\times$, which has order $p - 1$. The secret random numbers generated by Nikita and Michael are

$$n = 18319922375531859171613379181$$

and

$$m = 82335836243866695680141440300.$$

Nikita sends

$$g^n = 45416776270485369791375944998 \in (\mathbb{Z}/q)^\times$$

to Michael, and Michael sends

$$g^m = 15048074151770884271824225393 \in (\mathbb{Z}/q)^\times$$

to Nikita. They agree on the secret key

$$g^{nm} = 85771409470770521212346739540 \in (\mathbb{Z}/q)^\times.$$

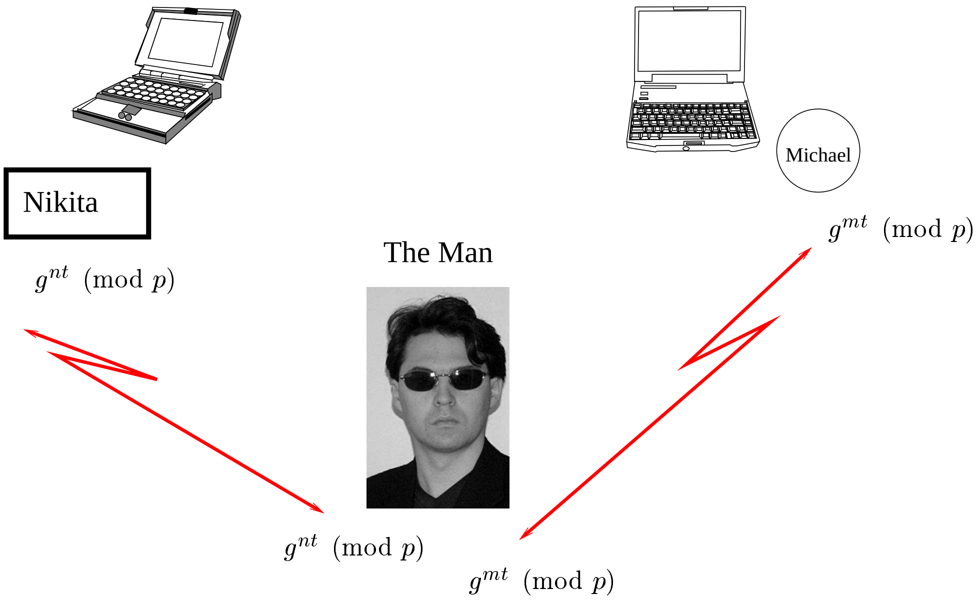


FIGURE 4.2. The Man in the Middle Attack

4.1.3 The Man in the Middle Attack

After their first system was broken, instead of talking on the phone, Michael and Nikita can now only communicate via text messages. Her captor, The Man, is watching each of the email transmissions; moreover, he can intercept messages and send false messages. When Nikita sends an email to Michael announcing $g^n \pmod p$, The Man intercepts this message, and sends his own number $g^t \pmod p$ to Michael. Eventually, Michael and The Man agree on the secret key $g^{tm} \pmod p$, and Nikita and The Man agree on the key $g^{tn} \pmod p$. When Nikita sends a message to Michael she foolishly uses the secret key $g^{tn} \pmod p$; The Man then intercepts it, decrypts it, changes it, and re-encrypts it using the key $g^{tm} \pmod p$, and sends it on to Michael. This is bad.

One way to get around this attack is to use "digital signatures" based on the RSA cryptosystem. We will not discuss digital signatures in this book, but we will discuss RSA in the next section.

4.2 The RSA Cryptosystem

The Diffie-Hellman key exchange has drawbacks. As discussed in Section 4.1.3, it is susceptible to the man in the middle attack, so one isn't always sure who they are exchanging messages with. Also, it only provides a way to agree on a secret key, not a way to encrypt any information; for that, one must rely on a symmetric-key encryption method. This section is about the RSA public-key cryptosystem of Rivest, Shamir, and Adleman [RSA78], which remedies some of these defects.

In this section we describe the RSA cryptosystem, then discuss several ways to attack it, which we must be aware of in order to implement the cryptosysem without making foolish mistakes.

4.2.1 How RSA works

The fundamental idea behind RSA is to try to construct a so-called "one-way function" on a set X, that is, an invertible function

$$E : X \to X$$

such that it is easy for Nikita to compute E^{-1}, but extremely difficult for anybody else to do so.

Here is how Nikita makes a one-way function E on the set of integers modulo n.

1. Nikita picks two large primes p and q, and lets $n = pq$.
2. It is easy for Nikita to then compute
$$\varphi(n) = \varphi(p) \cdot \varphi(q) = (p-1) \cdot (q-1).$$
3. Nikita next chooses a "random" integer e with
$$1 < e < \varphi(n) \text{ and } \gcd(e, \varphi(n)) = 1.$$
4. Nikita uses the algorithm from Section 3.5.2 to find a solution $x = d$ to the equation
$$ex \equiv 1 \pmod{\varphi(n)}.$$
5. Finally, Nikita defines a function $E : \mathbb{Z}/n \to \mathbb{Z}/n$ by
$$E(x) = x^e \in \mathbb{Z}/n.$$
Anybody can compute E fairly quickly using the repeated-squaring algorithm from Section 3.5.2.

Nikita's *public key* is the pair of integers (n, e), which is just enough information for people to easily compute E. Nikita knows a number d such that $ed \equiv 1 \pmod{\varphi(n)}$, so, as we will see below, she can quickly compute E^{-1}.

To send Nikita a message, proceed as follows. Encode your message, in some way, as a sequence of numbers modulo n (see Section 4.2.2)

$$m_1, \ldots, m_r \in \mathbb{Z}/n,$$

then send

$$E(m_1), \ldots, E(m_r)$$

to Nikita. (Recall that $E(m) = m^e$.)

When Nikita receives $E(m_i)$, she finds each m_i by using that $E^{-1}(m) = m^d$, a fact that follows from the following proposition.

Proposition 4.2.1. *Let n be an integer that is a product of distinct primes and let $d, e \in \mathbb{N}$ such that $p-1 \mid de-1$ for each prime $p \mid n$. Then $a^{de} \equiv a \pmod{n}$ for all $a \in \mathbb{Z}$.*

Proof. Since $n \mid a^{de} - a$ if and only if $p \mid a^{de} - a$ for each prime divisor p of n, it suffices to prove that $a^{de} \equiv a \pmod{p}$ for each prime divisor p of n. If $\gcd(a,p) \neq 0$, then $a \equiv 0 \pmod{p}$, so $a^{de} \equiv a \pmod{p}$. If $\gcd(a,p) = 1$, then Theorem 3.3.14 asserts that $a^{p-1} \equiv 1 \pmod{p}$. Since $p - 1 \mid de - 1$, we have $a^{de-1} \equiv 1 \pmod{p}$ as well. Multiplying both sides by a shows that $a^{de} \equiv a \pmod{p}$. $\square$

Thus to decrypt $E(m_i)$ Nikita computes

$$m_i = E^{-1}(E(m_i)) = E(m_i)^d = (m_i^e)^d = m_i.$$

4.2.2 Encoding a Phrase in a Number

In order to use the RSA cryptosystem to encrypt messages, it is necessary to encode them as a sequence of numbers of size less than $n = pq$. We now describe a simple way to do this.

Think of a sequence of capital letters and spaces as a number in base 27 as follows. Let a single space correspond to 0, the letter A to 1, B to 2, ..., Z to 26. Thus, e.g., "HARVARD" denotes a number written in base 27. The corresponding number written in decimal is 1808939906:

$$\begin{aligned}\text{HARVARD} &\leftrightarrow 8 + 27\cdot 1 + 27^2\cdot 18 + 27^3\cdot 22 + 27^4\cdot 1 + 27^5\cdot 18 + 27^6\cdot 4\\ &= 1808939906\end{aligned}$$

To recover the digits of the number, repeatedly divide by 27 and read off the remainder:

$$\begin{array}{rclcrl}
1808939906 &=& 66997774\cdot 27 &+& 8 & \text{"H"}\\
66997774 &=& 2481399\cdot 27 &+& 1 & \text{"A"}\\
2481399 &=& 91903\cdot 27 &+& 18 & \text{"R"}\\
91903 &=& 3403\cdot 27 &+& 22 & \text{"V"}\\
3403 &=& 126\cdot 27 &+& 1 & \text{"A"}\\
126 &=& 4\cdot 27 &+& 18 & \text{"R"}\\
4 &=& 0\cdot 27 &+& 4 & \text{"D"}
\end{array}$$

How many letters can a number hold? If $27^k < n$, then k letters can be encoded in a number $< n$. Put another way,

$$k < \log(n)/\log(27) = \log_{27}(n).$$

4.2.3 Examples

So the arithmetic is easy to follow, we use small primes p and q and encrypt the single letter "X".

1. Choose p and q: Let $p = 17$, $q = 19$, so $n = pq = 323$.
2. Compute $\varphi(n)$:
$$\begin{aligned}\varphi(n) &= \varphi(p \cdot q) = \varphi(p) \cdot \varphi(q) = (p-1)(q-1)\\ &= pq - p - q + 1 = 323 - 17 - 19 + 1 = 288.\end{aligned}$$
3. Randomly choose an $e < 288$: We choose $e = 95$.
4. Solve
$$95x \equiv 1 \pmod{288}.$$
Using the GCD algorithm, we find that $d = 191$ solves the equation.

The public key is $(323, 95)$, so the encryption function $E : \mathbb{Z}/323 \to \mathbb{Z}/323$ is defined by
$$E(x) = x^{95},$$
and the decryption function is $D(x) = x^{191}$.

Next, we encrypt the letter "X". It is encoded as the number 24, since X is the 24th letter of the alphabet. We have
$$E(24) = 24^{95} = 294 \in \mathbb{Z}/323.$$

To decrypt, we compute E^{-1}:
$$E^{-1}(294) = 294^{191} = 24 \in \mathbb{Z}/323.$$

This example illustrates RSA but with bigger numbers. Let
$$p = 738873402423833494183027176953, \; q = 3787776806865662882378273.$$
Then
$$n = p \cdot q = 2798687536910915970127263606347911460948554197853542169$$
and
$$\begin{aligned}\varphi(n) &= (p-1)(q-1)\\ &= 2798687536910915970127262867470721260308194351943986944.\end{aligned}$$
We somehow randomly chose
$$e = 1483959194866204179348536010284716655442139024915720699.$$
Then
$$d = 2113367928496305469541348387088632973457802358781610803$$

Since $\log_{27}(n)/\log(27) \approx 38.04$, we can encode then encrypt single blocks of up to 38 letters. Let's encrypt "HARVARD", which is encoded as $m = 1808939906$. We have

$$E(m) = 625425724974078486559370130768554070421628674916144724.$$

It is also interesting to note that changing the input message even slightly completely changes the encrypted version. For example, "HARVAHD" is encoded as $m' = 1665450836$ and encrypted as

$$E(m') = 437968760439188600589414766639328726464015666686231875.$$

4.3 Attacking RSA

Nikita's public key is (n, e). If we compute the factorization of $n = pq$, then we can compute $\varphi(n)$ and hence deduce her secret decoding number d. Thus attempting to factor n is a way to try to break an RSA public-key cryptosystem. In this lecture we consider several approaches to "cracking" RSA, and relate them to the difficulty of factoring n.

4.3.1 Factoring n Given $\varphi(n)$

Suppose $n = pq$. Given $\varphi(n)$, it is very easy to compute p and q. We have

$$\varphi(n) = (p-1)(q-1) = pq - (p+q) + 1,$$

so we know both $pq = n$ and $p + q = n + 1 - \varphi(n)$. Thus we know the polynomial

$$x^2 - (p+q)x + pq = (x-p)(x-q)$$

whose roots are p and q. These roots can be found using the quadratic formula.

Example 4.3.1. The number $n = pq = 31615577110997599711$ is a product of two primes, and $\varphi(n) = 31615577098574867424$. We have

$$\begin{aligned} f &= x^2 - (n + 1 - \varphi(n))x + n \\ &= x^2 - 12422732288x + 31615577110997599711 \\ &= (x - 3572144239)(x - 8850588049), \end{aligned}$$

where the last step is easily accomplished using the quadratic formula:

$$\begin{aligned} \frac{-b + \sqrt{b^2 - 4ac}}{2a} &= \frac{12422732288 + \sqrt{12422732288^2 - 4 \cdot 31615577110997599711}}{2} \\ &= 8850588049. \end{aligned}$$

We conclude that $n = 3572144239 \cdot 8850588049$.

4.3.2 When p and q are Close

Suppose that p and q are "close" to each other. Then it is easy to factor n using a factorization method of Fermat.

Suppose $n = pq$ with $p > q$, say. Then

$$n = \left(\frac{p+q}{2}\right)^2 - \left(\frac{p-q}{2}\right)^2.$$

Since p and q are "close",

$$s = \frac{p-q}{2}$$

is small,

$$t = \frac{p+q}{2}$$

is only slightly larger than $\sqrt{n}$, and $t^2 - n = s^2$ is a perfect square. So we just try

$$t = \lceil\sqrt{n}\rceil, \quad t = \lceil\sqrt{n}\rceil + 1, \quad t = \lceil\sqrt{n}\rceil + 2, \dots$$

until $t^2 - n$ is a perfect square s^2. (Here $\lceil x \rceil$ denotes the least integer $n \geq x$.) Then

$$p = t + s, \qquad q = t - s.$$

Example 4.3.2. Suppose $n = 23360947609$. Then

$$\sqrt{n} = 152842.88\ldots.$$

If $t = 152843$, then $\sqrt{t^2 - n} = 187.18\ldots$.
If $t = 152844$, then $\sqrt{t^2 - n} = 583.71\ldots$.
If $t = 152845$, then $\sqrt{t^2 - n} = 804 \in \mathbb{Z}$.
Thus $s = 804$. We find that $p = t + s = 153649$ and $q = t - s = 152041$.

4.3.3 Factoring n Given d

In this section, we show that cracking RSA is, in practice, at least as difficult as factoring n. We give a probabilistic algorithm that given a decryption key determines the factorization of n.

Suppose that we crack an RSA cryptosystem with modulus n and encryption key e by somehow finding an integer d such that

$$a^{ed} \equiv a \pmod{n}$$

for all a. Then $m = ed - 1$ satisfies $a^m \equiv 1 \pmod{n}$ for all a that are coprime to n. As we saw in Section 4.3.1, knowing $\varphi(n)$ leads directly to a factorization of n. Unfortunately, knowing d does not seem to lead easily to a factorization of n. However, there is a probabilistic procedure that, given an m such that $a^m \equiv 1 \pmod{n}$, will find a factorization of n with high probability.

Algorithm 4.3.3 (Probabilistic Algorithm to Factor n Given d). In the description of this algorithm, a always denotes an integer coprime to n. Given an integer $m > 1$ such that $a^m \equiv 1 \pmod{n}$ for all a, this probabilistic algorithm factors n with high probability.

1. *If* $a^{m/2} \equiv 1 \pmod{n}$ *for all* a, *replace* m *by* $m/2$. Note that m is even since $(-1)^m \equiv 1 \pmod{n}$. It is not practical to determine whether or not $a^{m/2} \equiv 1 \pmod{n}$ for all a, because it would require doing a computation for too many a. Instead, we try a few random a; if $a^{m/2} \equiv 1 \pmod{n}$ for the a we check, we divide m by 2.

   Note that if there exists even a single a such that $a^{m/2} \not\equiv 1 \pmod{n}$, then at least half the a have this property, since $a \mapsto a^{m/2}$ is a nontrivial homomorphism $(\mathbb{Z}/n)^\times \to \{\pm 1\}$ and the kernel can have size at most $\phi(n)/2 = \#(\mathbb{Z}/n)^\times/2$.

   Keep replacing m by $m/2$ until we find an a such that $a^{m/2} \not\equiv 1 \pmod{n}$.

2. *Try to factor* n *by computing gcd's.* Assume that we have found an m such that $a^m \equiv 1 \pmod{n}$ for all a coprime to n, but there is an a such that $a^{m/2} \not\equiv 1 \pmod{n}$. (That $x^2 \equiv 1 \pmod{p}$ implies $x = \pm 1 \pmod{p}$ follows from Proposition 5.1.1 in the next chapter.) Since $(a^{m/2})^2 \equiv 1 \pmod{n}$, we also have $(a^{m/2})^2 \equiv 1 \pmod{p}$ and $(a^{m/2})^2 \equiv 1 \pmod{q}$, so $a^{m/2} \equiv \pm 1 \pmod{p}$ and $a^{m/2} \equiv \pm 1 \pmod{q}$. Since $a^{m/2} \not\equiv 1 \pmod{n}$, their are three possibilities for these signs, so with probability 2/3,

   $$a^{m/2} \equiv +1 \pmod{p} \qquad \text{and} \qquad a^{m/2} \equiv -1 \pmod{q}$$

   or

   $$a^{m/2} \equiv -1 \pmod{p} \qquad \text{and} \qquad a^{m/2} \equiv +1 \pmod{q}.$$

   (The only other possibility is that both signs are -1.) In the first case,

   $$p \mid a^{m/2} - 1 \qquad \text{but} \qquad q \nmid a^{m/2} - 1,$$

   so $\gcd(a^{m/2} - 1, pq) = p$, and we have factored n. Similarly, in the second case, $\gcd(a^{m/2} - 1, pq) = q$, and we again factor n.

   Keep trying a's until one of these two cases occurs.

Example 4.3.4. Somehow we discover that the RSA cryptosystem with

$$n = 32295194023343 \qquad \text{and} \qquad e = 29468811804857$$

has decryption key $d = 11127763319273$. Let's use this information to factor n. We have

$$m = ed - 1 = 327921963064646896263108960.$$

For each $a \leq 20$ we find that $a^{m/2} \equiv 1 \pmod{n}$, so we replace m by

$$\frac{m}{2} = 163960981532323448131554480.$$

Again, we find with this new m that for each $a \leq 20$, $a^{m/2} \equiv 1 \pmod{n}$, so we replace m by 81980490766161724065777240. Yet again, for each $a \leq 20$,

$a^{m/2} \equiv 1 \pmod{n}$, so we replace m by 4099024538308086203288862O. This is enough, since $2^{m/2} \equiv 4015382800099 \pmod{n}$. Then

$$\gcd(2^{m/2} - 1, n) = \gcd(4015382800098, 32295194023343) = 737531,$$

and we have found a factor of n! Dividing, we find that

$$n = 737531 \cdot 43788253.$$

EXERCISES

4.1 You and Nikita wish to agree on a secret key using the Diffie-Hellman protocol. Nikita announces that $p = 3793$ and $g = 7$. Nikita secretely chooses a number $n < p$ and tells you that $g^n \equiv 454 \pmod{p}$. You choose the random number $m = 1208$. What is the secret key?

4.2 This problem concerns encoding phrases using numbers using the encoding of Section 4.2.2.

(a) Find the number that corresponds to `VE␣RI␣TAS`. (Note that the left-most "digit", `V`, is the least significant digit, and ␣ denotes a blank space.)

(b) What is the longest that an arbitrary sequence of letters (and space) can be if it must fit in a number that is less than 10^{20}?

4.3 You see Michael and Nikita agree on a secret key using the Diffie-Hellman key exchange protocol. Michael and Nikita choose $p = 97$ and $g = 5$. Nikita chooses a random number n and tells Michael that $g^n \equiv 3 \pmod{97}$, and Michael chooses a random number m and tells Nikita that $g^m \equiv 7 \pmod{97}$. Crack their code: What is the secret key that Nikita and Michael agree upon? What is n? What is m?

4.4 Using the RSA public key $(n, e) = (441484567519, 238402465195)$, encrypt the current year.

4.5 In this problem, you will "crack" an RSA cryptosystem.

(a) What is the secret decoding number d for the RSA cryptosystem with public key $(n, e) = (5352381469067, 4240501142039)$?

(b) The number 3539014000459 encrypts a question using the RSA cryptosystem from part (a). What is the question? (After decoding, you'll get a number. To find the corresponding word, see Section 4.2.2.)

4.6 Suppose Michael creates an RSA cryptosystem with a very large modulus n for which the factorization of n cannot be found in a reasonable amount of time. Suppose that Nikita sends messages to Michael by representing each alphabetic character as an integer between 0 and 26 (`A` corresponds to 1, `B` to 2, etc., and a space ␣ to 0), then encrypts each number *separately* using Michael's RSA cryptosystem. Is this method secure? Explain your answer.

4.7 Nikita creates an RSA cryptosystem with public key

$$(n, e) = (1433811615146881, 329222149569169).$$

In the following two problems, show the steps you take to factor n. (Don't simply factor n directly using a computer.)

(a) Somehow you discover that $d = 116439879930113$. Show how to use the probabilistic algorithm of Section 4.3.3 to use d to factor n.

(b) In part (a) you found that the factors p and q of n are very close. Show how to use the Fermat factorization method of Section 4.3.2 to factor n.

4.8 Nikita and Michael decide to agree on a secret encryption key using the Diffie-Hellman key exchange protocol. You observe the following:

(a) Nikita chooses $p = 13$ for the modulus and $g = 2$ as generator.

(b) Nikita sends 6 to Michael.

(c) Michael sends 11 to Nikita.

What is the secret key?

4.9 Consider the RSA public-key cryptosystem defined by $(n, e) = (77, 7)$.

(a) Encrypt the number 4 using this cryptosystem.

(b) Find an integer d such that $ed \equiv 1 \pmod{\varphi(n)}$.

4.10 Research the following: What is the current status of the RSA patent? Could you write a commercial program that implements the RSA cryptosystem without having to pay anyone royalties? What about a free program? Same questions, but for the Diffie-Hellman key exchange.

4.11 For any positive integer n, let $\sigma(n)$ be the sum of the divisors of n; for example, $\sigma(6) = 1+2+3+6 = 12$ and $\sigma(10) = 1+2+5+10 = 18$.

(a) (10 points) Suppose that $n = pqr$ with p, q, and r primes. Devise an "efficient" algorithm that given n, $\varphi(n)$ and $\sigma(n)$, computes the factorization of n. For example, if $n = 105$, then $p = 3$, $q = 5$, and $r = 7$, so the input to the algorithm would be

$$n = 105, \qquad \varphi(n) = 48, \qquad \text{and} \quad \sigma(n) = 192,$$

and the output would be 3, 5, 7.

(b) (3 points) Use your algorithm to factor $n = 60071026003$ given that $\varphi(n) = 60024000000$ and $\sigma(n) = 60118076016$.

5
The Structure of $(\mathbb{Z}/p)^\times$

This chapter is about the structure of the group $(\mathbb{Z}/p)^\times$ of units modulo p. The main result is that this group is always cyclic.

Definition 5.0.5 (Primitive root). A *primitive root* modulo an integer n is an element of $(\mathbb{Z}/n)^\times$ of order $\varphi(n)$.

We prove that there is a primitive root modulo every prime p. Since $(\mathbb{Z}/p)^\times$ has order $p-1$, this implies that $(\mathbb{Z}/p)^\times$ is a cyclic group, a fact this will be extremely useful, since it completely determines the structure of $(\mathbb{Z}/p)^\times$ as an abelian group.

If n is an odd prime power, then there is also a primitive root modulo n (see the exercises), but there is no primitive root modulo the even prime power 2^3.

Section 5.1 is the key input in our proof that $(\mathbb{Z}/p)^\times$ is cyclic; here we show that for every divisor d of $p-1$ there are exactly d elements of $(\mathbb{Z}/p)^\times$ whose order divides d. We then use this result in Section 5.2 to produce an element of $(\mathbb{Z}/p)^\times$ of order q^r when q^r is a prime power that exactly divides $p-1$ (i.e., q^r divides $p-1$, but q^{r+1} does not divide $p-1$), and combine together these to obtain an element of $(\mathbb{Z}/p)^\times$ of order $p-1$.

5.1 Polynomials over $\mathbb{Z}/p$

Proposition 5.1.1. *Let $f \in (\mathbb{Z}/p)[x]$ be a nonzero polynomial over the ring $\mathbb{Z}/p$. Then there are at most $\deg(f)$ elements $\alpha \in \mathbb{Z}/p$ such that $f(\alpha) = 0$.*

Proof. We induct on $\deg(f)$. The cases with $\deg(f) \leq 1$ are clear. Write $f = a_n x^n + \cdots a_1 x + a_0$. If $f(\alpha) = 0$ then

$$\begin{aligned}
f(x) &= f(x) - f(\alpha) \\
&= a_n(x^n - \alpha^n) + \cdots a_1(x - \alpha) + a_0(1 - 1) \\
&= (x - \alpha)(a_n(x^{n-1} + \cdots + \alpha^{n-1}) + \cdots + a_2(x + \alpha) + a_1) \\
&= (x - \alpha)g(x),
\end{aligned}$$

for some polynomial $g(x) \in (\mathbb{Z}/p)[x]$. Next suppose that $f(\beta) = 0$ with $\beta \neq \alpha$. Then $(\beta - \alpha)g(\beta) = 0$, so, since $\beta - \alpha \neq 0$, we have $g(\beta) = 0$. By our inductive hypothesis, g has at most $n - 1$ roots, so there are at most $n - 1$ possibilities for β. It follows that f has at most n roots. □

Proposition 5.1.2. *Let p be a prime number and let d be a divisor of $p - 1$. Then $f = x^d - 1 \in (\mathbb{Z}/p)[x]$ has exactly d solutions.*

Proof. Let $e = (p - 1)/d$. We have

$$\begin{aligned}
x^{p-1} - 1 &= (x^d)^e - 1 \\
&= (x^d - 1)((x^d)^{e-1} + (x^d)^{e-2} + \cdots + 1) \\
&= (x^d - 1)g(x),
\end{aligned}$$

where $g \in (\mathbb{Z}/p)[x]$ and $\deg(g) = de - d = p - 1 - d$. Theorem 3.3.14 implies that $x^{p-1} - 1$ has exactly $p - 1$ roots in $\mathbb{Z}/p$, since every nonzero element of $\mathbb{Z}/p$ is a root! By Proposition 5.1.1, g has *at most* $p - 1 - d$ roots and $x^d - 1$ has at most d roots. Since a root of $(x^d - 1)g(x)$ is a root of either $x^d - 1$ or $g(x)$ and $x^{p-1} - 1$ has $p - 1$ roots, g must have exactly $p - 1 - d$ roots and $x^d - 1$ must have exactly d roots, as claimed. □

The analogue of Proposition 5.1.2 is false when p is replaced by a composite integer n, since a root mod n of a product of two polynomials need not be a root of either factor. For example, if $n = n_1 \cdot n_2$ with $n_1, n_2 \neq 1$, then $f = n_1 x$ has at least *two* distinct zeros, namely 0 and $n_2 \neq 0$.

5.2 Existence of Primitive Roots

In this section, we prove that $(\mathbb{Z}/p)^\times$ is cylic by using the results of Section 5.2 to produce an element of $(\mathbb{Z}/p)^\times$ of order d for each prime power divisor d of $p - 1$, then multiply these together to obtain an element of order $p - 1$.

The following lemma will be used to assemble together elements of orders dividing $p - 1$ to produce an element of order $p - 1$.

Lemma 5.2.1. *Suppose $a, b \in (\mathbb{Z}/n)^\times$ have orders r and s, respectively, and that* $\gcd(r, s) = 1$. *Then ab has order rs.*

Proof. This is a general fact about commuting elements of any finite group. Since

$$(ab)^{rs} = a^{rs}b^{rs} = 1,$$

the order of ab is a divisor of rs. Write this divisor as $r_1 s_1$ where $r_1 \mid r$ and $s_1 \mid s$. Raise both sides of

$$a^{r_1 s_1} b^{r_1 s_1} = (ab)^{r_1 s_1} = 1.$$

to the power $r_2 = r/r_1$ to obtain

$$a^{r_1 r_2 s_1} b^{r_1 r_2 s_1} = 1.$$

Since $a^{r_1 r_2 s_1} = (a^{r_1 r_2})^{s_1} = 1$, we have

$$b^{r_1 r_2 s_1} = 1,$$

so $s \mid r_1 r_2 s_1$. Since $\gcd(s, r_1 r_2) = \gcd(s, r) = 1$, it follows that $s = s_1$. Similarly $r = r_1$, so the order of ab is rs. $\square$

Theorem 5.2.2. *There is a primitive root modulo any prime p.*

Proof. Write $p - 1$ as a product of distinct prime powers $q_i^{n_i}$:

$$p - 1 = q_1^{n_1} q_2^{n_2} \cdots q_r^{n_r}.$$

By Proposition 5.1.2, the polynomial $x^{q_i^{n_i}} - 1$ has exactly $q_i^{n_i}$ roots, and the polynomial $x^{q_i^{n_i - 1}} - 1$ has exactly $q_i^{n_i - 1}$ roots. There are $q_i^{n_i} - q_i^{n_i - 1} = q_i^{n_i - 1}(q_i - 1)$ elements $a \in \mathbb{Z}/p$ such that $a^{q_i^{n_i}} = 1$ but $a^{q_i^{n_i - 1}} \neq 1$; each of these elements has order $q_i^{n_i}$. Thus for each $i = 1, \ldots, r$, we can choose an a_i of order $q_i^{n_i}$. Then, using Lemma 5.2.1 repeatedly, we see that

$$a = a_1 a_2 \cdots a_r$$

has order $q_1^{n_1} \cdots q_r^{n_r} = p - 1$, so a is a primitive root modulo p. $\square$

Example 5.2.3. We illustrate the proof of Theorem 5.2.2 when $p = 13$. We have

$$p - 1 = 12 = 2^2 \cdot 3.$$

The polynomial $x^4 - 1$ has roots $\{1, 5, 8, 12\}$ and $x^2 - 1$ has roots $\{1, 12\}$, so we may take $a_1 = 5$. The polynomial $x^3 - 1$ has roots $\{1, 3, 9\}$, and we set $a_2 = 3$. Then $a = 5 \cdot 3 = 15 \equiv 2$ is a primitive root. To verify this, note that the successive powers of 2 modulo 13 are

$$2,\ 4,\ 8,\ 3,\ 6,\ 12,\ 11,\ 9,\ 5,\ 10,\ 7,\ 1.$$

Example 5.2.4. Theorem 5.2.2 is false if, e.g., p is replaced by a power of 2 bigger than 4. For example, the four elements of $(\mathbb{Z}/8)^\times$ each have order dividing 2, but $\varphi(8) = 4$.

Theorem 5.2.5. *Let p^n be a power of an odd prime. Then there is a primitive root modulo p^n.*

The proof is left as Exercise 3.

Proposition 5.2.6. *If there is a primitive root modulo n, then there are exactly $\varphi(\varphi(n))$ primitive roots modulo n.*

Proof. The primitive roots modulo n are the generators of $(\mathbb{Z}/n)^\times$, which by assumption is cyclic of order $\varphi(n)$. Thus they are in bijection with the generators of any cyclic group of order $\varphi(n)$. In particular, the number of primitive roots modulo n is the same as the number of elements of $\mathbb{Z}/\varphi(n)$ with additive order $\varphi(n)$. An element of $\mathbb{Z}/\varphi(n)$ has additive order $\varphi(n)$ if and only if it is coprime to $\varphi(n)$. There are $\varphi(\varphi(n))$ such elements, as claimed. □

For example, there are $\varphi(\varphi(17)) = \varphi(16) = 2^4 - 2^3 = 8$ primitive roots mod 17, namely $3, 5, 6, 7, 10, 11, 12, 14$. The $\varphi(\varphi(9)) = \varphi(6) = 2$ primitive roots modulo 9 are 2 and 5. There aren't any primitive roots modulo 8, even though $\varphi(\varphi(8)) = \varphi(4) = 2 > 0$.

5.3 Artin's Conjecture

Conjecture 5.3.1 (Emil Artin). *Suppose $a \in \mathbb{Z}$ is not -1 or a perfect square. Then there are infinitely many primes p such that a is a primitive root modulo p.*

There is no single integer a such that Artin's conjecture is known to be true. For any given a, Pieter [Mor93] proved that there are infinitely many p such that the order of a is divisible by the largest prime factor of $p - 1$.

Hooley [Hoo67] proved that the Generalized Riemann Hypothesis implies Conjecture 5.3.1. This Generalized Riemann Hypothesis is, as its name suggests, a generalization of the Riemann Hypothesis; it asserts that certain functions, called "zeta functions", have zeros only on the vertical line $\mathrm{Re}(s) = \frac{1}{2}$.

Remark 5.3.2. Artin conjectured more precisely that if $N(x, a)$ is the number of primes $p \le x$ such that a is a primitive root modulo p, then $N(x, a)$ is asymptotic to $C(a)\pi(x)$, where $C(a)$ is a positive constant that depends only on a and $\pi(x)$ is the number of primes up to x.

Exercises

5.1 Prove that there is no primitive root modulo 2^n for any $n \ge 3$. [Hint: Relate the statement for $n = 3$ to the statement for $n > 3$.]

5.2 Characterize the integers n such that there is a primitive root modulo n in terms of their prime factorization.

5.3 Let p be an odd prime.

   (a) Prove that there is a primitive root modulo p^2. [Hint: Write down an element of $(\mathbb{Z}/p^2)^\times$ that looks like it might have order p, and prove that it does. Recall that if a, b have orders n, m, with $\gcd(n, m) = 1$, then ab has order nm.]

   (b) Prove that for any n, there is a primitive root modulo p^n.

5.4 Search the literature for what is known about Artin's conjecture on primitive roots, and write a short survey of these results.

6 Quadratic Reciprocity

Let a be an integer. The quadratic reciprocity law of Gauss provides a beautiful and precise answer to the following question: "For which primes p is the image of a in $(\mathbb{Z}/p)^\times$ a perfect square?" Amazingly, the answer only depends on the residue of p modulo $4a$.

The quadratic reciprocity law has been proved in a huge number of ways (see [Lem] for a list). We give two distinct proofs. The first, which is elementary and involves tediously keeping track of integer points in intervals, is given Section 6.3. The second, given in Section 6.4, is extremely algebraic and uses congruences between sums of powers of the complex number $\zeta = e^{2\pi i/p}$. You should read Sections 6.1 and 6.2, then at least one of Section 6.3 or Section 6.4, depending on taste.

In Section 6.5, we return to the computational question of actually finding square roots in practice.

6.1 Statement of the Quadratic Reciprocity Law

In this section we motivate, then precisely state, the quadratic reciprocity law.

Definition 6.1.1 (Quadratic Residue). An integer a not divisible by a prime p is called a *quadratic residue* modulo p if a is a square modulo p. If a is not a square modulo p then a is called a *quadratic nonresidue*.

The quadratic reciprocity theorem connects the question of whether or not a is a quadratic residue modulo p to the question of whether p is a quadratic residue modulo each of the prime divisors of a. To express it precisely, we introduce some new notation. Let p be an odd prime and let a

TABLE 6.1. When is 5 a square modulo p?

p	$\left(\frac{5}{p}\right)$	$p \bmod 5$
7	-1	2
11	1	1
13	-1	3
17	-1	2
19	1	4
23	-1	3
29	1	4
31	1	1
37	-1	2
41	1	1
43	-1	3
47	-1	2

be an integer coprime to p. Set

$$\left(\frac{a}{p}\right) = \begin{cases} +1 & \text{if } a \text{ is a quadratic residue, and} \\ -1 & \text{otherwise.} \end{cases}$$

This notation is well entrenched in the literature, even though it is identical to the notation for "a divided by p"; be careful not to confuse the two.

Just as we defined $\gcd(a,b)$ for $a, b \in \mathbb{Z}/n$, define $\left(\frac{a}{p}\right)$ for $a \in \mathbb{Z}/p$ to be $\left(\frac{\tilde{a}}{p}\right)$ for any lift $\tilde{a}$ of a to $\mathbb{Z}$.

Proposition 6.2.1 below implies that

$$\left(\frac{a}{p}\right) \equiv a^{(p-1)/2} \pmod{p},$$

so the map $a \mapsto \left(\frac{a}{p}\right)$ is a multiplicative function in the sense that

$$\left(\frac{a}{p}\right) \cdot \left(\frac{b}{p}\right) = \left(\frac{ab}{p}\right).$$

The symbol $\left(\frac{a}{p}\right)$ only depends on the residue class of a modulo p. Thus tabulating the value of $\left(\frac{a}{5}\right)$ for hundreds of a would be silly, since it is so easy.

Question 6.1.2. Would it be equally silly to make a table of $\left(\frac{5}{p}\right)$ for many of primes p?

We find out by constructing Table 6.1 and looking for a simple pattern. It appears that $\left(\frac{5}{p}\right)$ depends only on the congruence class of p modulo 5. More precisely, $\left(\frac{5}{p}\right) = 1$ if and only if $p \equiv 1, 4 \pmod{5}$, i.e., $\left(\frac{5}{p}\right) = 1$ if and only if p is a square modulo 5. We might try to prove this using Proposition 6.2.1 below; however, I see no simple reason that knowing that

$p \equiv 1, 4 \pmod 5$ helps us to evaluate $5^{(p-1)/2} \pmod p$. See Exercise 4 for further a discussion about proving our observation directly.

Based on similar observations, in the 18th century various mathematicians found a conjectural explanation for the mystery suggested by Table 6.1. Finally, on April 8, 1796, at the age of only 19, Gauss proved the following theorem.

Theorem 6.1.3 (Quadratic Reciprocity Law). *Suppose that p and q are distinct odd primes. Then*

$$\left(\frac{p}{q}\right) = (-1)^{\frac{p-1}{2}\cdot\frac{q-1}{2}}\left(\frac{q}{p}\right).$$

Also

$$\left(\frac{-1}{p}\right) = (-1)^{(p-1)/2} \qquad \text{and} \qquad \left(\frac{2}{p}\right) = \begin{cases} 1 & \text{if } p \equiv \pm 1 \pmod 8 \\ -1 & \text{if } p \equiv \pm 3 \pmod 8. \end{cases}$$

We will give two proofs of Gauss's formula for $\left(\frac{p}{q}\right)$. The first very elementary proof is in Section 6.3, and the second more algebraic proof is in Section 6.4. The assertion about $\left(\frac{-1}{p}\right)$ will follow from Proposition 6.2.1 below. We only prove the assertion about $\left(\frac{2}{p}\right)$ in Section 6.3 (see Proposition 6.3.4), but do not give a corresponding proof in Section 6.4.

As expected, in our example Gauss's theorem implies that

$$\left(\frac{5}{p}\right) = (-1)^{2\cdot\frac{p-1}{2}}\left(\frac{p}{5}\right) = \left(\frac{p}{5}\right) = \begin{cases} +1 & \text{if } p \equiv 1, 4 \pmod 5 \\ -1 & \text{if } p \equiv 2, 3 \pmod 5. \end{cases}$$

The following example illustrates how to answer questions like "is a a square modulo b" using Theorem 6.1.3.

Example 6.1.4. Is 69 a square modulo 389? We have

$$\left(\frac{69}{389}\right) = \left(\frac{3\cdot 23}{389}\right) = \left(\frac{3}{389}\right)\cdot\left(\frac{23}{389}\right) = (-1)\cdot(-1) = 1.$$

Here

$$\left(\frac{3}{389}\right) = \left(\frac{389}{3}\right) = \left(\frac{2}{3}\right) = -1,$$

and

$$\begin{aligned}\left(\frac{23}{389}\right) &= \left(\frac{389}{23}\right) = \left(\frac{21}{23}\right) = \left(\frac{-2}{23}\right) \\ &= \left(\frac{-1}{23}\right)\left(\frac{2}{23}\right) = (-1)^{\frac{23-1}{2}}\cdot 1 = -1.\end{aligned}$$

Thus 69 is a square modulo 389.

Though we know that 69 is a square modulo 389, we don't know an explicit x such that $x^2 \equiv 6 \pmod{389}$! This is similar to how we could prove using Theorem 3.3.14 that certain numbers are composite without knowing a factorization, except that it is easy in practice to find square roots, as we'll discuss in Section 6.5 and Example 6.5.1.

6.2 Euler's Criterion

Let p be an odd prime and a an integer not divisible by p. Euler used the existence of primitive roots to show that $\left(\frac{a}{p}\right)$ is congruent to $a^{(p-1)/2}$ modulo p. We will use this fact repeatedly below in both proofs of Theorem 6.1.3.

Proposition 6.2.1 (Euler's Criterion). *Then* $\left(\frac{a}{p}\right) = 1$ *if and only if*

$$a^{(p-1)/2} \equiv 1 \pmod{p}.$$

Proof. By Theorem 5.2.2, there is an integer g that has order $p-1$ modulo p, so every integer coprime to p is congruent to a power of g. First suppose that a is congruent to a perfect square modulo p, so

$$a \equiv (g^r)^2 \equiv g^{2r} \pmod{p}$$

for some r. Then by Theorem 3.3.14

$$a^{(p-1)/2} \equiv g^{2r \cdot \frac{p-1}{2}} \equiv g^{r(p-1)} \equiv 1 \pmod{p}.$$

Conversely, suppose that $a^{(p-1)/2} \equiv 1 \pmod{p}$. We have $a \equiv g^r \pmod{p}$ for some integer r. Thus $g^{r(p-1)/2} \equiv 1 \pmod{p}$, so

$$p-1 \mid r(p-1)/2$$

which implies that r is even. Thus $a \equiv (g^{r/2})^2 \pmod{p}$, so a is congruent to a square modulo p. $\square$

Corollary 6.2.2. *The equation* $x^2 \equiv a \pmod{p}$ *has no solution if and only if* $a^{(p-1)/2} \equiv -1 \pmod{p}$. *Thus* $\left(\frac{a}{p}\right) \equiv a^{(p-1)/2} \pmod{p}$.

Proof. This follows from Proposition 6.2.1 and the fact that the polynomial $x^2 - 1$ has no roots besides $+1$ and -1 (which follows from Proposition 5.1.2). $\square$

Example 6.2.3. Suppose $p = 11$. By squaring each element of $(\mathbb{Z}/11)^\times$, we see that the squares modulo 11 are $\{1, 3, 4, 5, 9\}$. We compute $a^{(p-1)/2} = a^5$ for each $a \in (\mathbb{Z}/11)^\times$ and get

$$1^5 = 1,\ 2^5 = -1,\ 3^5 = 1,\ 4^5 = 1,\ 5^5 = 1,$$
$$6^5 = -1,\ 7^5 = -1,\ 8^5 = -1,\ 9^5 = 1,\ 10^5 = -1.$$

Thus the a with $a^5 = 1$ are $\{1, 3, 4, 5, 9\}$, just as Proposition 6.2.1 predicts.

Example 6.2.4. We determine whether or not 3 is a square modulo the prime $p = 726377359$. Using a computer we find that

$$3^{(p-1)/2} \equiv -1 \pmod{726377359}.$$

Thus 3 is not a square modulo p. This computation wasn't difficult, but it would have been tedious by hand. The law of quadratic reciprocity provides a way to answer this questions that could easily be carried out by hand:

$$\begin{aligned}\left(\frac{3}{726377359}\right) &= (-1)^{(3-1)/2\cdot(726377359-1)/2}\left(\frac{726377359}{3}\right)\\ &= (-1)\cdot\left(\frac{1}{3}\right) = -1.\end{aligned}$$

It is a general fact that if G is any abelian group and n is any integer, then the map $x \mapsto x^n$ is a homomorphism. Thus, in group-theoretic language, Proposition 6.2.1 asserts that the map

$$\left(\frac{\bullet}{p}\right) : (\mathbb{Z}/p)^\times \to \{\pm 1\}$$

that sends a to $\left(\frac{a}{p}\right)$ is a homomorphism of groups.

Proposition 6.2.5. *The homomorphism* $\left(\frac{\bullet}{p}\right) : (\mathbb{Z}/p)^\times \to \{\pm 1\}$ *is surjective.*

Proof. If $\left(\frac{\bullet}{p}\right)$ is not surjective, then $\left(\frac{a}{p}\right) = 1$ for every $a \in (\mathbb{Z}/p)^\times$. This means that the squaring map $a \mapsto a^2$ on $(\mathbb{Z}/p)^\times$ is surjective. But -1 is in the kernel of squaring and $(\mathbb{Z}/p)^\times$ is finite, so squaring is not surjective. □

6.3 First Proof of Quadratic Reciprocity

Our first proof of quadratic reciprocity is elementary but tedious. The proof involves keeping track of integer points in intervals. Proving Gauss's lemma is the first step; this lemma computes $\left(\frac{a}{p}\right)$ in terms of the number of integers of a certain type that lie in a certain interval. Next we prove Lemma 6.3.2, which controls how the parity of the number of integer points in an interval changes when an endpoint of the interval is changed. Then we prove that $\left(\frac{a}{p}\right)$ only depends on p modulo $4a$ by applying Gauss's lemma and keeping careful track of intervals as they are rescaled and their endpoints changed. Finally, in Section 6.3.2 we use some basic algebra to deduce the quadratic reciprocity law using the tools we've just developed.

Lemma 6.3.1 (Gauss's Lemma). *Let p be an odd prime and let a be an integer $\not\equiv 0 \pmod{p}$. Form the numbers*

$$a,\ 2a,\ 3a,\ \ldots,\ \frac{p-1}{2}a$$

and reduce them modulo p to lie in the interval $(-\frac{p}{2},\ \frac{p}{2})$. Let ν be the number of negative numbers in the resulting set. Then

$$\left(\frac{a}{p}\right) = (-1)^\nu.$$

Proof. In defining ν, we expressed each number in

$$S = \left\{a, 2a, \ldots, \frac{p-1}{2}a\right\}$$

as congruent to a number in the set

$$\left\{1, -1, 2, -2, \ldots, \frac{p-1}{2}, -\frac{p-1}{2}\right\}.$$

No number $1, 2, \ldots, \frac{p-1}{2}$ appears more than once, with either choice of sign, because if it did then either two elements of S are congruent modulo p or 0 is the sum of two elements of S, and both events are impossible. Thus the resulting set must be of the form

$$T = \left\{\varepsilon_1 \cdot 1, \varepsilon_2 \cdot 2, \ldots, \varepsilon_{(p-1)/2} \cdot \frac{p-1}{2}\right\},$$

where each ε_i is either $+1$ or -1. Multiplying together the elements of S and of T, we see that

$$(1a) \cdot (2a) \cdot (3a) \cdots \left(\frac{p-1}{2}a\right) \equiv$$
$$(\varepsilon_1 \cdot 1) \cdot (\varepsilon_2 \cdot 2) \cdots \left(\varepsilon_{(p-1)/2} \cdot \frac{p-1}{2}\right) \pmod{p},$$

so

$$a^{(p-1)/2} \equiv \varepsilon_1 \cdot \varepsilon_2 \cdots \varepsilon_{(p-1)/2} \pmod{p}.$$

The lemma then follows from Proposition 6.2.1, since $\left(\frac{a}{p}\right) = a^{(p-1)/2}$. □

6.3.1 Euler's Conjecture

For rational numbers $a, b \in \mathbb{Q}$, let

$$(a, b) \cap \mathbb{Z} = \{x \in \mathbb{Z} : a \leq x \leq b\}$$

be the set of integers between a and b. The following Lemma will help us to keep track of how many integers lie in certain intervals.

Lemma 6.3.2. *Let $a, b \in \mathbb{Q}$. Then for any integer n,*

$$\#\left((a, b) \cap \mathbb{Z}\right) \equiv \#\left((a, b + 2n) \cap \mathbb{Z}\right) \pmod{2}$$

and

$$\#\left((a, b) \cap \mathbb{Z}\right) \equiv \#\left((a - 2n, b) \cap \mathbb{Z}\right) \pmod{2},$$

provided that each interval involved in the congruence is nonempty.

The statement is illustrated in Figure 6.1. Note that if one of the intervals is empty, then the statement is false; e.g., if $(a, b) = (-1/2, 1/2)$ and $n = -1$ then $\#((a, b) \cap \mathbb{Z}) = 1$ but $\#(a, b - 2) \cap \mathbb{Z} = 0$.

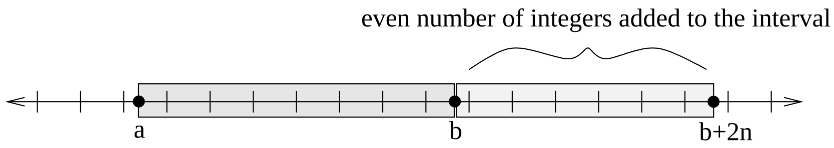


FIGURE 6.1. Illustration of Lemma 6.3.2

Proof. Since $n > 0$,

$$(a, b + 2n) = (a, b) \cup [b, b + 2n),$$

where the union is disjoint. Recall that $\lceil x \rceil$ denotes the least integer $\geq x$. There are $2n$ integers,

$$\lceil b \rceil, \lceil b \rceil + 1, \ldots, \lceil b \rceil + 2n - 1,$$

in the interval $[b, b + 2n)$, so the first congruence of the lemma is true in this case. We also have

$$(a, b - 2n) = (a, b) \backslash [b - 2n, b)$$

and $[b - 2n, b)$ also contains exactly $2n$ integers, so the lemma is also true when n is negative. The statement about $\#\left((a - 2n, b) \cap \mathbb{Z}\right)$ is proved in a similar manner. $\square$

The following proposition was conjectured by Euler, based on extensive numerical evidence. Once we've proved this proposition, it will be easy to deduce the quadratic reciprocity law.

Proposition 6.3.3 (Euler's Conjecture). *Let p be an odd prime and a a positive integer with $p \nmid a$.*

1. *The symbol $\left(\frac{a}{p}\right)$ depends only on p modulo $4a$.*

2. *If q is a prime with $q \equiv -p \pmod{4a}$, then $\left(\frac{a}{p}\right) = \left(\frac{a}{q}\right)$.*

Proof. We will apply Lemma 6.3.1 to compute $\left(\frac{a}{p}\right)$. Let

$$S = \left\{a, 2a, 3a, \ldots, \frac{p-1}{2}a\right\}$$

and

$$I = \left(\frac{1}{2}p, p\right) \cup \left(\frac{3}{2}p, 2p\right) \cup \cdots \cup \left(\left(b - \frac{1}{2}\right)p, bp\right),$$

where $b = \frac{1}{2}a$ or $\frac{1}{2}(a-1)$, whichever is an integer. We check that every element of S that reduces to something in the interval $(-\frac{p}{2}, 0)$ lies in I. This is clear if $b = \frac{1}{2}a < \frac{p-1}{2}a$. If $b = \frac{1}{2}(a-1)$, then $bp + \frac{p}{2} > \frac{p-1}{2}a$, so $((b - \frac{1}{2})p, bp)$ is the last interval that could contain an element of S that reduces to $(-\frac{p}{2}, 0)$. Note that the integer endpoints of I are not in S, since those endpoints are divisible by p, but no element of S is divisible by p. Thus, by Lemma 6.3.1,

$$\left(\frac{a}{p}\right) = (-1)^{\#(S\cap I)}.$$

To compute $\#(S\cap I)$, first rescale by a to see that

$$\#(S\cap I) = \#\left(\mathbb{Z}\cap \frac{1}{a}I\right),$$

where

$$\frac{1}{a}I = \left(\left(\frac{p}{2a}, \frac{p}{a}\right) \cup \left(\frac{3p}{2a}, \frac{2p}{a}\right) \cup \cdots \cup \left(\frac{(2b-1)p}{2a}, \frac{bp}{a}\right)\right).$$

Write $p = 4ac + r$, and let

$$J = \left(\left(\frac{r}{2a}, \frac{r}{a}\right) \cup \left(\frac{3r}{2a}, \frac{2r}{a}\right) \cup \cdots \cup \left(\frac{(2b-1)r}{2a}, \frac{br}{a}\right)\right).$$

The only difference between I and J is that the endpoints of intervals are changed by addition of an even integer. By Lemma 6.3.2,

$$\nu = \#\left(\mathbb{Z}\cap \frac{1}{a}I\right) \equiv \#(\mathbb{Z}\cap J) \pmod{2}.$$

Thus $\left(\frac{a}{p}\right) = (-1)^\nu$ depends only on r, i.e., only on p modulo $4a$.

If $q \equiv -p \pmod{4a}$, then the only change in the above computation is that r is replaced by $4a - r$. This changes $\frac{1}{a}I$ into

$$\begin{aligned} K = \left(2 - \frac{r}{2a}, 4 - \frac{r}{a}\right) \cup \left(6 - \frac{3r}{2a}, 8 - \frac{2r}{a}\right) \cup \cdots \\ \cup \left(4b - 2 - \frac{(2b-1)r}{2a}, 4b - \frac{br}{a}\right). \end{aligned}$$

Thus K is the same as $-\frac{1}{a}I$, except even integers have been added to the endpoints. By Lemma 6.3.2,

$$\#(K\cap\mathbb{Z}) \equiv \#\left(\left(\frac{1}{a}I\right)\cap\mathbb{Z}\right) \pmod{2},$$

so $\left(\frac{a}{p}\right) = \left(\frac{a}{q}\right)$, which completes the proof. □

The following more careful analysis in the special case when $a = 2$ helps illustrate the proof of the above lemma, and is frequently useful in computations.

Proposition 6.3.4. *Let p be an odd prime. Then*

$$\left(\frac{2}{p}\right) = \begin{cases} 1 & \text{if } p \equiv \pm 1 \pmod{8} \\ -1 & \text{if } p \equiv \pm 3 \pmod{8}. \end{cases}$$

Proof. When $a = 2$, the set $S = \{a, 2a, \ldots, 2 \cdot \frac{p-1}{2}\}$ is

$$\{2, 4, 6, \ldots, p-1\}.$$

We must count the parity of the number of elements of S that lie in the interval $I = (\frac{p}{2}, p)$. Writing $p = 8c + r$, we have

$$\begin{aligned}\#\,(I \cap S) &= \#\left(\frac{1}{2}I \cap \mathbb{Z}\right) = \#\left(\left(\frac{p}{4}, \frac{p}{2}\right) \cap \mathbb{Z}\right)\\ &= \#\left(\left(2c + \frac{r}{4}, 4c + \frac{r}{2}\right) \cap \mathbb{Z}\right) \equiv \#\left(\left(\frac{r}{4}, \frac{r}{2}\right) \cap \mathbb{Z}\right) \pmod{2},\end{aligned}$$

where the last equality comes from Lemma 6.3.2. The possibilities for r are $1, 3, 5, 7$. When $r = 1$, the cardinality is 0, when $r = 3, 5$ it is 1, and when $r = 7$ it is 2. $\square$

6.3.2 *Proof of Quadratic Reciprocity*

It is now straightforward to deduce the quadratic reciprocity law.

First Proof of Theorem 6.1.3. First suppose that $p \equiv q \pmod{4}$. By swapping p and q if necessary, we may assume that $p > q$, and write $p - q = 4a$. Since $p = 4a + q$,

$$\left(\frac{p}{q}\right) = \left(\frac{4a+q}{q}\right) = \left(\frac{4a}{q}\right) = \left(\frac{4}{q}\right)\left(\frac{a}{q}\right) = \left(\frac{a}{q}\right),$$

and

$$\left(\frac{q}{p}\right) = \left(\frac{p-4a}{p}\right) = \left(\frac{-4a}{p}\right) = \left(\frac{-1}{p}\right) \cdot \left(\frac{a}{p}\right).$$

Proposition 6.3.3 implies that $\left(\frac{a}{q}\right) = \left(\frac{a}{p}\right)$, since $p \equiv q \pmod{4a}$. Thus

$$\left(\frac{p}{q}\right) \cdot \left(\frac{q}{p}\right) = \left(\frac{-1}{p}\right) = (-1)^{\frac{p-1}{2}} = (-1)^{\frac{p-1}{2} \cdot \frac{q-1}{2}},$$

where the last equality is because $\frac{p-1}{2}$ is even if and only if $\frac{q-1}{2}$ is even.

Next suppose that $p \not\equiv q \pmod{4}$, so $p \equiv -q \pmod{4}$. Write $p + q = 4a$. We have

$$\left(\frac{p}{q}\right) = \left(\frac{4a-q}{q}\right) = \left(\frac{a}{q}\right), \quad \text{and} \quad \left(\frac{q}{p}\right) = \left(\frac{4a-p}{p}\right) = \left(\frac{a}{p}\right).$$

Since $p \equiv -q \pmod{4a}$, Proposition 6.3.3 implies that $\left(\frac{p}{q}\right) = \left(\frac{q}{p}\right)$. Since $(-1)^{\frac{p-1}{2} \cdot \frac{q-1}{2}} = 1$, the proof is complete. $\square$

Example 6.3.5. Is 3 a square modulo $p = 726377359$? We proved that the answer is "no" in the previous lecture by computing $3^{p-1} \pmod{p}$. It's easier to prove that the answer is no using Theorem 6.1.3:

$$\left(\frac{3}{726377359}\right) = (-1)^{1 \cdot \frac{726377358}{2}} \cdot \left(\frac{726377359}{3}\right) = -\left(\frac{1}{3}\right) = -1.$$

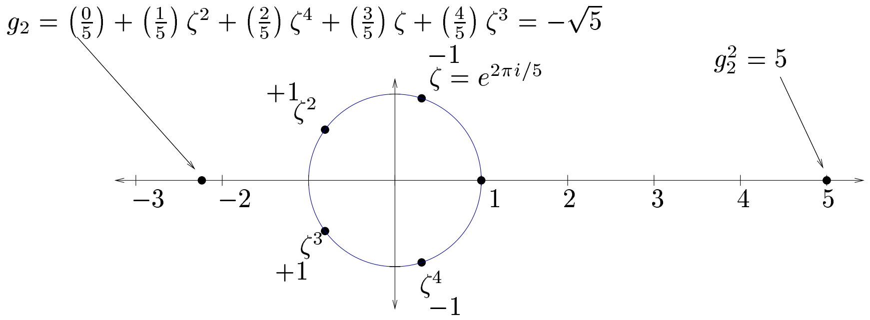


FIGURE 6.2. Gauss sum g_2 for $p = 5$

6.4 A Proof of Quadratic Reciprocity Using Gauss Sums

In this section we present a beautiful proof of Theorem 6.1.3 using algebraic identities satisfied by sums of "roots of unity". The objects we introduce in the proof are of independent interest, and provide a powerful tool to prove higher-degree analogues of quadratic reciprocity. (For more on higher reciprocity see [IR90]. See also Section 6 of [IR90] on which the proof below is modeled.)

Recall that a *complex number* is a number of the form $a+b\sqrt{-1}$, where a and b are real numbers, and that the set of complex numbers forms a field.

Definition 6.4.1 (Root of Unity). An nth *root of unity* is a complex number ζ such that $\zeta^n = 1$. A root of unity is a *primitive* nth root of unity if n is the smallest positive integer such that $\zeta^n = 1$.

Since for θ a real number, $e^{i\theta} = \cos(\theta) + i\sin(\theta)$, the complex number $e^{2\pi i/n}$ is a primitive nth root unity. For the rest of this section, fix a prime p and a primitive pth root ζ of unity, e.g., $\zeta = e^{2\pi i/p}$.

Definition 6.4.2 (Gauss Sum). The *Gauss sum* associated to an integer a is

$$g_a = \sum_{n=0}^{p-1} \left(\frac{n}{p}\right) \zeta^{an}.$$

(Note that p is implicit in the definition of g_a. If we were to change p, then the Gauss sum g_a associated to a would be different.)

Figure 6.2 illustrates the Gauss sum g_2 for $p = 5$. The Gauss sum is got by adding the points on the unit circle, with signs as indicated, to obtain the real number $-\sqrt{5}$. This suggests the following proposition, whose proof will require some work.

Proposition 6.4.3. *For any a not divisible by p,*

$$g_a^2 = (-1)^{(p-1)/2} p.$$

In order to prove the proposition, we introduce a few lemmas.

Lemma 6.4.4. *For any integer a,*

$$\sum_{n=0}^{p-1} \zeta^{an} = \begin{cases} p, & \text{if } a \equiv 0 \pmod{p} \\ 0, & \text{otherwise.} \end{cases}$$

Proof. If $a \equiv 0 \pmod{p}$, then $\zeta^a = 1$, so the sum equals the number of summands, which is p. If $a \not\equiv 0 \pmod{p}$, we use the telescopic identity $x^p - 1 = (x-1)(x^{p-1} + \cdots + x + 1)$ with $x = \zeta^a$. We have $\zeta^a \neq 1$, so $\zeta^a - 1 \neq 0$ and

$$\sum_{n=0}^{p-1} \zeta^{an} = \frac{\zeta^{ap} - 1}{\zeta^a - 1} = 0.$$

$\square$

Lemma 6.4.5. *Let x and y be integers and let $\delta(x,y)$ be 1 if $x \equiv y \pmod{p}$ or 0 otherwise. Then*

$$\sum_{n=0}^{p-1} \zeta^{(x-y)n} = p \cdot \delta(x,y).$$

Proof. This follows immediately from Lemma 6.4.4 by setting $a = x - y$. $\square$

Lemma 6.4.6. *Let p be a prime. Then*

$$g_0 = \sum_{n=0}^{p-1} \left(\frac{n}{p}\right) = 0.$$

Proof. By Proposition 6.2.5, the map

$$\left(\frac{\bullet}{p}\right) : (\mathbb{Z}/p)^\times \to \{\pm 1\}$$

is a surjective homomorphism of groups. Thus exactly half the elements of $(\mathbb{Z}/p)^\times$ map to $+1$ and half map to -1 (the subgroup that maps to $+1$ has index 2). Since $\left(\frac{0}{p}\right) = 0$, the sum in the statement of the lemma is 0. $\square$

Lemma 6.4.7. *Let p be a prime and a any integer. Then*

$$g_a = \left(\frac{a}{p}\right) g_1.$$

Proof. When $a \equiv 0 \pmod{p}$ the lemma follows immediately from Lemma 6.4.6, so suppose that $a \not\equiv 0 \pmod{p}$. Then

$$\left(\frac{a}{p}\right) g_a = \left(\frac{a}{p}\right) \sum_{n=0}^{p-1} \left(\frac{n}{p}\right) \zeta^{an} = \sum_{n=0}^{p-1} \left(\frac{an}{p}\right) \zeta^{an} = \sum_{m=0}^{p-1} \left(\frac{m}{p}\right) \zeta^{m} = g_1.$$

Now multiply both sides by $\left(\frac{a}{p}\right)$ and use that $\left(\frac{a}{p}\right)^2 = 1$. $\square$

We now have enough lemmas to prove Proposition 6.4.3.

Proof of Proposition 6.4.3. We evaluate the sum $\sum_{a=0}^{p-1} g_a g_{-a}$ in two different ways. By Lemma 6.4.7, since $a \not\equiv 0 \pmod{p}$ we have

$$g_a g_{-a} = \left(\frac{a}{p}\right) g_1 \left(\frac{-a}{p}\right) g_1 = \left(\frac{-1}{p}\right)\left(\frac{a}{p}\right)^2 g_1^2 = (-1)^{(p-1)/2} g_1^2,$$

where the last step follows from Proposition 6.2.1 and the fact that $\left(\frac{a}{p}\right) \in \{\pm 1\}$. Thus

$$\sum_{a=0}^{p-1} g_a g_{-a} = (p-1)(-1)^{(p-1)/2} g_1^2. \tag{6.1}$$

On the other hand, by definition

$$\begin{aligned} g_a g_{-a} &= \sum_{n=0}^{p-1} \left(\frac{n}{p}\right) \zeta^{an} \cdot \sum_{m=0}^{p-1} \left(\frac{m}{p}\right) \zeta^{-am} \\ &= \sum_{n=0}^{p-1}\sum_{m=0}^{p-1} \left(\frac{n}{p}\right)\left(\frac{m}{p}\right) \zeta^{an}\zeta^{-am} \\ &= \sum_{n=0}^{p-1}\sum_{m=0}^{p-1} \left(\frac{n}{p}\right)\left(\frac{m}{p}\right) \zeta^{an-am}. \end{aligned}$$

Thus by Lemma 6.4.5,

$$\begin{aligned} \sum_{a=0}^{p-1} g_a g_{-a} &= \sum_{a=0}^{p-1}\sum_{n=0}^{p-1}\sum_{m=0}^{p-1} \left(\frac{n}{p}\right)\left(\frac{m}{p}\right) \zeta^{an-am} \\ &= \sum_{n=0}^{p-1}\sum_{m=0}^{p-1} \left(\frac{n}{p}\right)\left(\frac{m}{p}\right) \sum_{a=0}^{p-1} \zeta^{an-am} \\ &= \sum_{n=0}^{p-1}\sum_{m=0}^{p-1} \left(\frac{n}{p}\right)\left(\frac{m}{p}\right) p\delta(n,m) \\ &= \sum_{n=0}^{p-1} \left(\frac{n}{p}\right)^2 p = p(p-1). \end{aligned}$$

Equating (6.1) and the above equality then canceling $(p-1)$ shows that

$$g_1^2 = (-1)^{(p-1)/2} p.$$

Since $a \not\equiv 0 \pmod{p}$, we have $\left(\frac{a}{p}\right)^2 = 1$, so by Lemma 6.4.7,

$$g_a^2 = \left(\frac{a}{p}\right)^2 g_1^2 = g_1^2,$$

and the proposition is proved. $\square$

6.4.1 Proof of Quadratic Reciprocity

We are now in a position to prove Theorem 6.1.3 using Gauss sums.

Proof. Let q be an odd prime with $q \neq p$. Set $p^* = (-1)^{(p-1)/2}p$ and recall that Proposition 6.4.3 asserts that $p^* = g^2$, where $g = g_1 = \sum_{n=0}^{p-1} \left(\frac{n}{p}\right) \zeta^n$ is a Gauss sum with $\zeta = e^{2\pi i/p}$ a primitive pth root of unity.

Proposition 6.2.1 trivially implies that

$$(p^*)^{(q-1)/2} \equiv \left(\frac{p^*}{q}\right) \pmod{q}.$$

We have $g^{q-1} = (g^2)^{(q-1)/2} = (p^*)^{(q-1)/2}$, so multiplying both sides of the displayed equation by g yields a congruence

$$g^q \equiv g\left(\frac{p^*}{q}\right) \pmod{q}. \tag{6.2}$$

But what does this congruence *mean*, given that g^q is not an integer? In Exercise 8, you will prove that every $\mathbb{Z}$-linear combination of powers of ζ can be written uniquely as a $\mathbb{Z}$-linear combination of elements of $B = \{1, \zeta, \ldots, \zeta^{p-2}\}$. The above congruence means that if we write g^q and $g\left(\frac{p^*}{q}\right)$ as $\mathbb{Z}$-linear combinations of the elements of B then the coefficients of the linear combination are congruent modulo q.

Another useful property of congruences, which you will prove in Exercise 9, is that if x and y are two $\mathbb{Z}$-linear combinations of powers of ζ, then $(x+y)^q \equiv x^q + y^q \pmod{q}$. Applying this, we see that

$$g^q = \left(\sum_{n=0}^{p-1} \left(\frac{n}{p}\right) \zeta^n\right)^q \equiv \sum_{n=0}^{p-1} \left(\frac{n}{p}\right)^q \zeta^{nq} \equiv \sum_{n=0}^{p-1} \left(\frac{n}{p}\right) \zeta^{nq} \equiv g_q \pmod{q}.$$

By Lemma 6.4.7,

$$g^q \equiv g_q \equiv \left(\frac{q}{p}\right) g \pmod{q}.$$

Combining this with (6.2) yields

$$\left(\frac{q}{p}\right) g \equiv \left(\frac{p^*}{q}\right) g \pmod{q}.$$

Since $g^2 = p^*$ and $p \neq q$, we can cancel g from both sides to find that $\left(\frac{q}{p}\right) \equiv \left(\frac{p^*}{q}\right) \pmod{q}$. Since both residue symbols are ± 1 and q is odd, it follows that $\left(\frac{q}{p}\right) = \left(\frac{p^*}{q}\right)$. Finally, we note using Proposition 6.2.1 that

$$\left(\frac{p^*}{q}\right) = \left(\frac{(-1)^{(p-1)/2}p}{q}\right) = \left(\frac{-1}{q}\right)^{(p-1)/2} \left(\frac{p}{q}\right) = (-1)^{\frac{q-1}{2}\cdot\frac{p-1}{2}} \cdot \left(\frac{p}{q}\right).$$

$\square$

6.5 How To Find Square Roots

After all this theory, we return in this section to the computational question of computing square roots.

One of the first things a school child learns in their algebra course is that the solutions to the quadratic equation

$$ax^2 + bx + c = 0 \qquad\qquad (\text{with } a \neq 0)$$

are

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

as one can see, e.g., by substituting the formula on the right back into the quadratic equation.

In school, $a \neq 0$, b, and c are typically chosen to be real or complex numbers. We're grown up now, so let p be an odd prime, and suppose instead that $a, b, c \in \mathbb{Z}/p$. Then the quadratic formula still gives solutions to $ax^2 + bx + c = 0$, and using Proposition 5.1.1 we can see that it gives all of them.

Using Theorem 6.1.3, we can decide whether or not $b^2 - 4ac$ is a perfect square, and hence whether or not $ax^2 + bx + c = 0$ has a solution in $\mathbb{Z}/p$. If $b^2 - 4ac$ *is* a perfect square, Theorem 6.1.3 says nothing about finding an actual square root. Also, note that for this problem we do *not* need quadratic reciprocity; in practice to decide whether an element of $\mathbb{Z}/p$ is a perfect square Proposition 6.2.1 is fast, in light of Section 3.5.

Suppose $a \in \mathbb{Z}/p$ is a nonzero quadratic residue. If $p \equiv 3 \pmod 4$ then $b = a^{\frac{p+1}{4}}$ is a square root of a because

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

There is no publically known deterministic polynomial time algorithm to compute a square root of a when $p \equiv 1 \pmod 4$. The following is a probabilistic algorithm to compute a square root of a. Let R be the ring $(\mathbb{Z}/p)[x]/(x^2 - a)$. Thus

$$R = \{u + vx : u, v \in \mathbb{Z}/p\}$$

with

$$(u + vx)(z + wx) = (uz + awv) + (uw + vz)x.$$

Let b and c be the square roots of a (we can't compute b and c at this stage, but we can consider them in order to deduce an algorithm to find them). Then by a generalization of the Chinese Remainder Theorem, there is a ring isomorphism

$$\varphi : R \longrightarrow \mathbb{Z}/p \times \mathbb{Z}/p$$

given by $\varphi(u + vx) = (u + vb, u + vc)$. Let z be a random element of $(\mathbb{Z}/p)^\times$ and let $u + vx = (1 + zx)^{\frac{p-1}{2}}$. If $v \neq 0$ we can quickly find b and c as follows. The quantity $u + vb$ is a $(p-1)/2$th power in $\mathbb{Z}/p$, so it equals either 0, 1, or -1. Thus $b = -u/v$, $(1-u)/v$, or $(-1-u)/v$. Since we know u and v we can try each of $-u/v$, $(1-u)/v$, and $(-1-u)/v$ and see which is a square root of a.

Example 6.5.1. Continuing example 6.1.4, we find a square root of 69 modulo 389. We apply the algorithm described above in the case $p \equiv 1 \pmod 4$. We first choose the random element $1+24x$, and find that $(1+24x)^{194} = -1$. The coefficient of x in the power is 0, so we try again. This time we have $(1+51x)^{194} = 239x = u + vx$. The inverse of 239 in $\mathbb{Z}/389$ is 153, so we consider the following three possibilities for a square root of 69:

$$-\frac{u}{v} = 0 \qquad \frac{1-u}{v} = 153 \qquad -\frac{1-u}{v} = -153.$$

Thus 153 and -153 are the square roots of 69 in $\mathbb{Z}/389$.

Exercises

6.1 Calculate the following symbols by hand: $\left(\frac{3}{97}\right)$, $\left(\frac{5}{389}\right)$, $\left(\frac{2003}{11}\right)$, and $\left(\frac{5!}{7}\right)$.

6.2 Prove that for $p \geq 5$ prime, $\left(\frac{3}{p}\right) = \begin{cases} 1 & \text{if } p \equiv 1, 11 \pmod{12}, \\ -1 & \text{if } p \equiv 5, 7 \pmod{12}. \end{cases}$

6.3 Use the fact that $(\mathbb{Z}/p)^\times$ is cyclic to give a direct proof that $\left(\frac{-3}{p}\right) = 1$ when $p \equiv 1 \pmod 3$. [Hint: There is an $c \in (\mathbb{Z}/p)^\times$ of order 3. Show that $(2c+1)^2 = -3$.]

6.4 If $p \equiv 1 \pmod 5$, show directly that $\left(\frac{5}{p}\right) = 1$ by the method of Exercise 3. [Hint: Let $c \in (\mathbb{Z}/p)^\times$ be an element of order 5. Show that $(c+c^4)^2 + (c+c^4) - 1 = 0$, etc.]

6.5 For which primes p is $\displaystyle\sum_{a=1}^{p-1} \left(\frac{a}{p}\right) = 0$?

6.6 How many natural numbers $x < 2^{13}$ satisfy the equation

$$x^2 \equiv 5 \pmod{2^{13}-1}?$$

(You may assume that $2^{13}-1$ is prime.)

6.7 Find the natural number $x < 97$ such that $x \equiv 4^{48} \pmod{97}$. (Note that 97 is prime.)

6.8 Let p be a prime and let ζ be a primitive pth root of unity. Prove that every $\mathbb{Z}$-linear combination of powers of ζ can be written uniquely as a $\mathbb{Z}$-linear combination of elements of $B = \{1, \zeta, \ldots, \zeta^{p-2}\}$. [Hint: $\zeta^p - 1 = 0$, so $\zeta^{p-1} + \cdots + \zeta + 1 = 0$, so $\zeta^{p-1} = -(\zeta^{p-2} + \cdots + \zeta + 1)$. Next prove that the polynomial $x^{p-1} + \cdots + x + 1$ does not factor over $\mathbb{Q}$.]

6.9 Let p be a prime and let ζ be a primitive pth root of unity. Suppose that x and y are $\mathbb{Z}$-linear combinations of powers of ζ. Prove that $(x+y)^p \equiv x^p + y^p \pmod p$.

6.10 Formulate an analogue of quadratic reciprocity for $\left(\frac{a}{q}\right)$ but without the restriction that q be a prime. By "analogue of quadratic reciprocity", I mean an easy way to tell whether or not a is a square modulo q. [Hint: Use Theorem 3.4.2 to reduce to the case where q is a prime power. Prove that if p is an odd prime that doesn't divide a then a is a square modulo p if and only if a is a square modulo p^n for any positive n.]

7
Continued Fractions

A *continued fraction* is an expression of the form

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}}$$

where the a_i are real numbers and $a_i > 0$ for $i \geq 1$; the expression may or may not go on indefinitely. We denote the value of this continued fraction by

$$[a_0, a_1, a_2, \ldots].$$

For example,

$$[1, 2] = 1 + \frac{1}{2} = \frac{3}{2},$$

and

$$\frac{172}{51} = [3, 2, 1, 2, 6] = 3 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{6}}}}.$$

Continued fractions have many applications, from the abstract to the concrete. For example, they are useful in understanding Pell's equation $x^2 - dy^2 = 1$, they give good rational approximations to irrational numbers, and provide a superb way to recognize a decimal approximation to a rational number. Continued fractions also suggest a sense in which e might be "less transcendental" than π (see Example 7.2.3).

There are many places to read about continued fractions, including [HW79, Ch. X], [Bur89, §13.3], and [Khi63]. This chapter was probably most influenced by Hardy and Wright.

In Section 7.1 we study continued fractions $[a_0, a_1, \ldots, a_n]$ of finite length and lay foundations for our later investigations. In Section 7.2 we give the continued fraction algorithm, which associates to a real number x a sequence $a_0, a_1, \ldots$ of integers such that $x = \lim_{n\to\infty}[a_0, a_1, \ldots, a_n]$. We also prove that if $a_0, a_1, \ldots$ is any infinite sequence of positive integers, then the sequence $c_n = [a_0, a_1, \ldots, a_n]$ converges; more generally, we prove that if the a_n are arbitrary real numbers and $\sum_{n=0}^{\infty} a_n$ diverges then (c_n) converges. In Section 7.3, we prove that a continued fraction with $a_i \in \mathbb{Z}$ is (eventually) periodic if and only if its value is a nonrational root of a quadratic polynomial, then discuss our extreme ignorance about continued fractions of roots of irreducible polynomials of degree greater than 2. In Section 7.4 we conclude the chapter with applications of continued fractions to recognizing approximations to rational numbers and solving Pell's equation ($x^2 - dy^2 = 1$).

7.1 Finite Continued Fractions

This section is about continued fractions of finite length. The main ideas are a recursive definition of numbers p_n and q_n such that

$$[a_0, a_1, \ldots, a_n] = \frac{p_n}{q_n},$$

and a formula for the determinants of $\left(\begin{smallmatrix} p_n & p_{n-1} \\ q_n & q_{n-1} \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} p_n & p_{n-2} \\ q_n & q_{n-2} \end{smallmatrix}\right)$. We will use the determinant formula to deduce properties of the sequence of partial convergents $[a_0, \ldots, a_k]$, and the Euclidean algorithm to prove that every rational number is represented by a continued fraction.

Definition 7.1.1. A *finite continued fraction* is an expression

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\cdots + \frac{1}{a_n}}}},$$

where each a_m is a real number and $a_m > 0$ for all $m \geq 1$. If the a_m are all integers, we say that the continued fraction is *integral.*

To get a feeling for continued fractions, observe that

$$\begin{aligned}
[a_0] &= a_0, \\
[a_0, a_1] &= a_0 + \frac{1}{a_1} = \frac{a_0 a_1 + 1}{a_1}, \\
[a_0, a_1, a_2] &= a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2}} = \frac{a_0 a_1 a_2 + a_0 + a_2}{a_1 a_2 + 1}.
\end{aligned}$$

Also,

$$\begin{aligned}[a_0, a_1, \dots, a_{m-1}, a_m] &= [a_0, a_1, \dots, a_{m-2}, a_{m-1} + \frac{1}{a_m}] \\ &= a_0 + \frac{1}{[a_1, \dots, a_m]} \\ &= [a_0, [a_1, \dots, a_m]].\end{aligned}$$

7.1.1 Partial Convergents

Fix a continued fraction $[a_0, \dots, a_m]$.

Definition 7.1.2. For $0 \leq n \leq m$, the nth *convergent* of the continued fraction $[a_0, \dots, a_m]$ is $[a_0, \dots, a_n]$.

For each $n \geq -1$, define real numbers p_n and q_n as follows:

$$\begin{array}{lllll} p_{-2} = 0, & p_{-1} = 1, & p_0 = a_0, & \cdots & p_n = a_n p_{n-1} + p_{n-2}, \\ q_{-2} = 1, & q_{-1} = 0, & q_0 = 1, & \cdots & q_n = a_n q_{n-1} + q_{n-2}. \end{array}$$

Proposition 7.1.3. $[a_0, \dots, a_n] = \dfrac{p_n}{q_n}$.

Proof. We use induction. We already verified the assertion when $n = 0, 1$. Suppose the proposition is true for all continued fractions of length $n - 1$. Then

$$\begin{aligned}[a_0, \dots, a_n] &= [a_0, \dots, a_{n-2}, a_{n-1} + \frac{1}{a_n}] \\ &= \frac{\left(a_{n-1} + \frac{1}{a_n}\right) p_{n-2} + p_{n-3}}{\left(a_{n-1} + \frac{1}{a_n}\right) q_{n-2} + q_{n-3}} \\ &= \frac{(a_{n-1}a_n + 1)p_{n-2} + a_n p_{n-3}}{(a_{n-1}a_n + 1)q_{n-2} + a_n q_{n-3}} \\ &= \frac{a_n(a_{n-1}p_{n-2} + p_{n-3}) + p_{n-2}}{a_n(a_{n-1}q_{n-2} + q_{n-3}) + q_{n-2}} \\ &= \frac{a_n p_{n-1} + p_{n-2}}{a_n q_{n-1} + q_{n-2}} = \frac{p_n}{q_n}.\end{aligned}$$

□

Proposition 7.1.4. *Suppose* $n \leq m$.

1. *The determinant of* $\begin{pmatrix} p_n & p_{n-1} \\ q_n & q_{n-1} \end{pmatrix}$ *is* $(-1)^{n-1}$*; equivalently,*
$$\frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}} = (-1)^{n-1} \cdot \frac{1}{q_n q_{n-1}}.$$
2. *The determinant of* $\begin{pmatrix} p_n & p_{n-2} \\ q_n & q_{n-2} \end{pmatrix}$ *is* $(-1)^n a_n$*; equivalently,*
$$\frac{p_n}{q_n} - \frac{p_{n-2}}{q_{n-2}} = (-1)^n \cdot \frac{a_n}{q_n q_{n-2}}.$$

Proof. For the first statement, we proceed by induction. The case $n = 0$ holds because the determinant of $\left(\begin{smallmatrix} a_0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ is $-1 = (-1)^{-1}$. Suppose the statement is true for $n - 1$. Then

$$\begin{aligned} p_n q_{n-1} - q_n p_{n-1} &= (a_n p_{n-1} + p_{n-2}) q_{n-1} - (a_n q_{n-1} + q_{n-2}) p_{n-1} \\ &= p_{n-2} q_{n-1} - q_{n-2} p_{n-1} \\ &= -(p_{n-1} q_{n-2} - p_{n-2} q_{n-1}) \\ &= -(-1)^{n-2} = (-1)^{n-1}. \end{aligned}$$

This completes the proof of the first statement. For the second statement,

$$\begin{aligned} p_n q_{n-2} - p_{n-2} q_n &= (a_n p_{n-1} + p_{n-2}) q_{n-2} - p_{n-2}(a_n q_{n-1} + q_{n-2}) \\ &= a_n (p_{n-1} q_{n-2} - p_{n-2} q_{n-1}) \\ &= (-1)^n a_n. \end{aligned}$$

□

Corollary 7.1.5. *The fraction* $\dfrac{p_n}{q_n}$ *is in lowest terms.*

Proof. If $p \mid p_n$ and $p \mid q_n$ then $p \mid (-1)^{n-1}$. □

7.1.2 How the Convergents Converge

Let $[a_0, \ldots, a_m]$ be a continued fraction and for $n \leq m$ let

$$c_n = [a_0, \ldots, a_n] = \frac{p_n}{q_n}$$

denote the nth convergent.

Proposition 7.1.6. *The even convergents c_{2n} increase strictly with n, and the odd convergents c_{2n+1} decrease strictly with n. Moreover, the odd convergents c_{2n+1} are greater than all of the even convergents.*

Proof. For $n \geq 1$ the a_n are positive, so the q_n are all positive. By Proposition 7.1.4, for $n \geq 2$,

$$c_n - c_{n-2} = (-1)^n \cdot \frac{a_n}{q_n q_{n-2}},$$

which proves the first claim.

Next, Proposition 7.1.4 implies that for $n \geq 1$,

$$c_n - c_{n-1} = (-1)^{n-1} \cdot \frac{1}{q_n q_{n-1}}$$

has the sign of $(-1)^{n-1}$, so that $c_{2n+1} > c_{2n}$. Thus if there exists r, n such that $c_{2n+1} < c_{2r}$, then $r \neq n$. If $r < n$, then $c_{2n+1} < c_{2r} < c_{2n}$, a contradiction. If $r > n$, then $c_{2r+1} < c_{2n+1} < c_{2r}$, also a contradiction. □

7.1.3 Every Rational Number is Represented

Proposition 7.1.7. *Every rational number is represented by a continued fraction.*

Proof. Without loss of generality we may assume that the rational number is a/b, with $b > 1$ and $\gcd(a, b) = 1$. Euclid's algorithm gives:

$$\begin{aligned}
a &= b \cdot a_0 + r_1, & 0 < r_1 < b\\
b &= r_1 \cdot a_1 + r_2, & 0 < r_2 < r_1\\
&\cdots\\
r_{n-2} &= r_{n-1} \cdot a_{n-1} + r_n, & 0 < r_n < r_{n-1}\\
r_{n-1} &= r_n \cdot a_n + 0.
\end{aligned}$$

Note that $a_i > 0$ for $i > 0$ (also $r_n = 1$ since $\gcd(a, b) = 1$). Rewrite the equations as follows:

$$\begin{aligned}
a/b &= a_0 + r_1/b = a_0 + 1/(b/r_1),\\
b/r_1 &= a_1 + r_2/r_1 = a_1 + 1/(r_1/r_2),\\
r_1/r_2 &= a_2 + r_3/r_2 = a_2 + 1/(r_2/r_3),\\
&\cdots\\
r_{n-1}/r_n &= a_n.
\end{aligned}$$

It follows that

$$\frac{a}{b} = [a_0, a_1, \ldots a_n].$$

□

The representation of a rational number as a continued fraction is not unique. For example, $2 = [1, 1] = [2]$.

7.2 Infinite Continued Fractions

This section begins with the continued fraction algorithm, which associates to a real number x a sequence $a_0, a_1, \ldots$ of integers. After giving several examples, we prove that $x = \lim_{n\to\infty}[a_0, a_1, \ldots, a_n]$ by proving that the odd and even partial convergents become arbitrarily close to each other. We also show that if $a_0, a_1, \ldots$ is any infinite sequence of positive integers, then the sequence of $c_n = [a_0, a_1, \ldots, a_n]$ converges, and, more generally, if a_n is an arbitrary sequence such that $\sum_{n=0}^{\infty} a_n$ diverges then (c_n) converges.

7.2.1 The Continued Fraction Algorithm

Let $x \in \mathbb{R}$ and write

$$x = a_0 + t_0$$

with $a_0 \in \mathbb{Z}$ and $0 \le t_0 < 1$. If $t_0 \ne 0$, write

$$\frac{1}{t_0} = a_1 + t_1$$

with $a_1 \in \mathbb{N}$ and $0 \le t_1 < 1$. Thus $t_0 = \frac{1}{a_1+t_1} = [0, a_1 + t_1]$, which is a (nonintegral) continued fraction expansion of t_0. Continue in this manner so long as $t_n \neq 0$ writing

$$\frac{1}{t_n} = a_{n+1} + t_{n+1}$$

with $a_{n+1} \in \mathbb{N}$ and $0 \le t_{n+1} < 1$. This process, which associates to a real number x the sequence of integers $a_0, a_1, a_2, \ldots$, is called the *continued fraction algorithm.*

Example 7.2.1. Let $x = \frac{8}{3}$. Then $x = 2 + \frac{2}{3}$, so $a_0 = 2$ and $t_0 = \frac{2}{3}$. Then $\frac{1}{t_0} = \frac{3}{2} = 1 + \frac{1}{2}$, so $a_1 = 1$ and $t_1 = \frac{1}{2}$. Then $\frac{1}{t_1} = 2$, so $a_2 = 2$, $t_2 = 0$, and the sequence terminates. Notice that

$$\frac{8}{3} = [2, 1, 2],$$

so the continued fraction algorithm produces the continued fraction of $\frac{8}{3}$.

Example 7.2.2. Let $x = \frac{1+\sqrt{5}}{2}$. Then

$$x = 1 + \frac{-1+\sqrt{5}}{2},$$

so $a_0 = 1$ and $t_0 = \frac{-1+\sqrt{5}}{2}$. We have

$$\frac{1}{t_0} = \frac{2}{-1+\sqrt{5}} = \frac{-2-2\sqrt{5}}{-4} = \frac{1+\sqrt{5}}{2}$$

so again $a_1 = 1$ and $t_1 = \frac{-1+\sqrt{5}}{2}$. Likewise, $a_n = 1$ for all n. As we will see below, the following crazy-looking equality makes sense.

$$\frac{1+\sqrt{5}}{2} = 1 + \cfrac{1}{1 + \cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cdots}}}}}$$

Example 7.2.3. Suppose $x = e = 2.71828182\ldots$. Applying the continued fraction algorithm, we have

$$a_0, a_1, a_2, \ldots = 2, 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, 1, 1, 10, \ldots$$

We have

$$[a_0, a_1, a_2, a_3, a_4, a_5] = \frac{87}{32} = 2.71875$$

which is a good rational approximation to e.

Let's do the same thing with $\pi = 3.14159265358979\ldots$: We have

$$a_0, a_1, a_2, \ldots = 3, 7, 15, 1, 292, 1, 1, 1, 2, 1, 3, 1, 14, \ldots$$

The first few partial convergents are

$$3, \frac{22}{7}, \frac{333}{106}, \frac{355}{113}, \frac{103993}{33102}, \cdots$$

These are all good rational approximations to π; for example,

$$\frac{103993}{33102} = 3.14159265301\ldots.$$

Notice that the continued fraction of e appears to exhibit a nice pattern, whereas the continued fraction of π appears fairly random. In some vague sense, this suggests that π is "more transcendental" than e. (That the continued fraction of e is as suggested by our above computation can be proved, though this is not easy; see, e.g., [Per50, §31].)

7.2.2 Convergence of Infinite Continued Fractions

Lemma 7.2.4. *For every n such that a_n is defined, we have*

$$x = [a_0, a_1, \ldots, a_n + t_n],$$

and if $t_n \neq 0$ then $x = [a_0, a_1, \ldots, a_n, \frac{1}{t_n}]$.

Proof. Use induction. The statements are both true when $n = 0$. If the second statement is true for $n-1$, then

$$\begin{aligned} x &= [a_0, a_1, \ldots, a_{n-1}, \frac{1}{t_{n-1}}] \\ &= [a_0, a_1, \ldots, a_{n-1}, a_n + t_n] \\ &= [a_0, a_1, \ldots, a_{n-1}, a_n, \frac{1}{t_n}]. \end{aligned}$$

Similarly, the first statement is true for n if it is true for $n-1$. □

Theorem 7.2.5. *Let $a_0, a_1, a_2, \ldots$ be a sequence of integers such that $a_n > 0$ for all $n \geq 1$, and for each $n \geq 0$, set $c_n = [a_0, a_1, \ldots a_n]$. Then $\lim_{n\to\infty} c_n$ exists.*

Proof. For any $m \geq n$, the number c_n is a partial convergent of $[a_0, \ldots, a_m]$. By Proposition 7.1.6 the even convergents c_{2n} form a strictly *increasing* sequence and the odd convergents c_{2n+1} form a strictly *decreasing* sequence. Moreover, the even convergents are all $\leq c_1$ and the odd convergents are all $\geq c_0$. Hence $\alpha_0 = \lim_{n\to\infty} c_{2n}$ and $\alpha_1 = \lim_{n\to\infty} c_{2n+1}$ both exist and $\alpha_0 \leq \alpha_1$. Finally, by Proposition 7.1.4

$$|c_{2n} - c_{2n-1}| = \frac{1}{q_{2n} \cdot q_{2n-1}} \leq \frac{1}{2n(2n-1)} \to 0,$$

so $\alpha_0 = \alpha_1$. □

We define

$$[a_0, a_1, \ldots] = \lim_{n\to\infty} c_n.$$

Example 7.2.6. We illustrate the theorem with $x = \pi$. As in the proof of Theorem 7.2.5, let c_n be the nth partial convergent to π. The c_n with n odd converge down to π

$$c_1 = 3.1428571\ldots, \; c_3 = 3.1415929\ldots, \; c_5 = 3.1415926\ldots$$

whereas the c_n with n even converge up to π

$$c_2 = 3.1415094\ldots,\ c_4 = 3.1415926\ldots,\ c_6 = 3.1415926\ldots.$$

Theorem 7.2.7. *Let $a_0, a_1, a_2, \ldots$ be a sequence of real numbers such that $a_n > 0$ for all $n \geq 1$, and for each $n \geq 0$, set $c_n = [a_0, a_1, \ldots a_n]$. Then $\lim_{n\to\infty} c_n$ exists if and only if the sum $\sum_{n=0}^{\infty} a_n$ diverges.*

Proof. We only prove that if $\sum a_n$ diverges then $\lim_{n\to\infty} c_n$ exists. A proof of the converse can be found in [Wal48, Ch. 2, Thm. 6.1].

Let q_n be the sequence of "denominators" of the partial convergents, as defined in Section 7.1.1, so $q_{-2} = 1$, $q_{-1} = 0$, and for $n \geq 0$,

$$q_n = a_n q_{n-1} + q_{n-2}.$$

As we saw in the proof of Theorem 7.2.5, the limit $\lim_{n\to\infty} c_n$ exists provided that the sequence $(q_n q_{n-1})$ diverges to infinity, in the sense that for every M there exists N for which $q_n q_{n-1} > M$ for all $n > N$.

For n even,

$$\begin{aligned}
q_n &= a_n q_{n-1} + q_{n-2} \\
&= a_n q_{n-1} + a_{n-2} q_{n-3} + q_{n-4} \\
&= a_n q_{n-1} + a_{n-2} q_{n-3} + a_{n-4} q_{n-5} + q_{n-6} \\
&= a_n q_{n-1} + a_{n-2} q_{n-3} + \cdots + a_2 q_1 + q_0
\end{aligned}$$

and for n odd,

$$q_n = a_n q_{n-1} + a_{n-2} q_{n-3} + \cdots + a_1 q_0 + q_{-1}.$$

Since $a_n > 0$ for $n > 0$, the sequence (q_n) is increasing; also $q_0 = a_0 q_{-1} + q_{-2} = 1$. Thus $q_i \geq 1$ for all $i \geq 0$. Applying this fact to the above expressions for q_n, we see that for n even

$$q_n \geq a_n + a_{n-2} + \cdots + a_2,$$

and for n odd

$$q_n \geq a_n + a_{n-2} + \cdots + a_1.$$

If $\sum a_n$ diverges, then at least one of $\sum a_{2n}$ or $\sum a_{2n+1}$ must diverge. The above inequalities then imply that at least one of the sequences (q_{2n}) or (q_{2n+1}) diverge to infinity. Since (q_n) is an increasing sequence, it follows that $(q_n q_{n-1})$ diverges to infinity. □

Example 7.2.8. Let $a_n = \frac{1}{n \log(n)}$ for $n \geq 2$ and $a_0 = a_1 = 0$. By the integral test, $\sum a_n$ diverges, so by Theorem 7.2.7 the continued fraction $[a_0, a_1, a_2, \ldots]$ converges. This convergence is very slow, since e.g.

$$[a_0, a_1, \ldots, a_{9999}] = 0.5750039671012225425930\ldots$$

yet

$$[a_0, a_1, \ldots, a_{10000}] = 0.7169153932917378550424\ldots.$$

Theorem 7.2.9. *Let $x \in \mathbb{R}$ be a real number. Then*

$$x = [a_0, a_1, a_2, \ldots],$$

where $a_0, a_1, a_2, \ldots$ is the sequence produced by the continued fraction algorithm.

Proof. If the sequence is finite then some $t_n = 0$ and the result follows by Lemma 7.2.4. Suppose the sequence is infinite. By Lemma 7.2.4,

$$x = [a_0, a_1, \ldots, a_n, \frac{1}{t_n}].$$

By Proposition 7.1.3 (which we apply in a case when the partial quotients of the continued fraction are not integers!), we have

$$x = \frac{\frac{1}{t_n} \cdot p_n + p_{n-1}}{\frac{1}{t_n} \cdot q_n + q_{n-1}}.$$

Thus if $c_n = [a_0, a_1, \ldots, a_n]$, then

$$\begin{aligned} x - c_n &= x - \frac{p_n}{q_n} \\ &= \frac{\frac{1}{t_n} p_n q_n + p_{n-1} q_n - \frac{1}{t_n} p_n q_n - p_n q_{n-1}}{q_n \left(\frac{1}{t_n} q_n + q_{n-1}\right)}. \\ &= \frac{p_{n-1} q_n - p_n q_{n-1}}{q_n \left(\frac{1}{t_n} q_n + q_{n-1}\right)} \\ &= \frac{(-1)^n}{q_n \left(\frac{1}{t_n} q_n + q_{n-1}\right)}. \end{aligned}$$

Thus

$$\begin{aligned} |x - c_n| &= \frac{1}{q_n \left(\frac{1}{t_n} q_n + q_{n-1}\right)} \\ &< \frac{1}{q_n (a_{n+1} q_n + q_{n-1})} \\ &= \frac{1}{q_n \cdot q_{n+1}} \leq \frac{1}{n(n+1)} \to 0. \end{aligned}$$

(In the inequality we use that a_{n+1} is the integer part of $\frac{1}{t_n}$, and is hence $\leq \frac{1}{t_n} < 1$, since $t_n < 1$.)

□

The following corollary follows from the proof of the above theorem.

Corollary 7.2.10. *Let $a_0, a_1, \ldots$ define an integral continued fraction, and let $x = [a_0, a_1, \ldots] \in \mathbb{R}$ be its value. Then for all m,*

$$\left| x - \frac{p_m}{q_m} \right| < \frac{1}{q_m \cdot q_{m+1}}.$$

Proposition 7.2.11. *If x is a rational number then the sequence $a_0, a_1, a_2, \ldots$ produced by the continued fraction algorithm terminates.*

Proof. Let $[b_0, b_1, \ldots, b_m]$ be the continued fraction representation of x that we obtain using the Euclidean algorithm. Then

$$x = b_0 + 1/[b_1, \ldots, b_m].$$

If $[b_1, \ldots, b_m] = 1$ then $m = 1$ and $b_1 = 1$, which will not happen using the Euclidean algorithm, since it would give $[b_0 + 1]$ for the continued fraction of the integer $b_0 + 1$. Thus $[b_1, \ldots, b_m] > 1$, so in the continued fraction algorithm we choose $a_0 = b_0$ and $t_0 = 1/[b_1, \ldots, b_m]$. Repeating this argument enough times proves the claim. □

7.3 Quadratic Irrationals

The main result of this section is that the continued fraction expansion of a number is eventually repeating if and only if the number is a quadratic irrational. This can be viewed as an analogue for continued fractions of the familiar fact that the decimal expansion of x is eventually repeating if and only if x is rational. The proof that continued fractions of quadratic irrationals eventually repeats is surprisingly difficult and involves an interesting finiteness argument. Section 7.3.3 emphasizes our striking ignorance about continued fractions of real roots of irreducible polynomials over $\mathbb{Q}$ of degree bigger than 2.

7.3.1 Quadratic Irrationals

Definition 7.3.1. An element $\alpha \in \mathbb{R}$ is a *quadratic irrational* if it is irrational and satisfies a quadratic polynomial with coefficients in $\mathbb{Q}$.

Thus, e.g., $(1+\sqrt{5})/2$ is a quadratic irrational. Recall that

$$\frac{1+\sqrt{5}}{2} = [1, 1, 1, \ldots].$$

The continued fraction of $\sqrt{2}$ is $[1, 2, 2, 2, 2, 2, \ldots]$, and the continued fraction of $\sqrt{389}$ is

$$[19, 1, 2, 1, 1, 1, 1, 2, 1, 38, 1, 2, 1, 1, 1, 1, 2, 1, 38, \ldots].$$

Does the $[1, 2, 1, 1, 1, 1, 2, 1, 38]$ pattern repeat over and over again?

7.3.2 Periodic Continued Fractions

Definition 7.3.2. A *periodic continued fraction* is a continued fraction $[a_0, a_1, \ldots, a_n, \ldots]$ such that

$$a_n = a_{n+h}$$

for a fixed positive integer h and all sufficiently large n. We call h the *period* of the continued fraction.

Example 7.3.3. Consider the periodic continued fraction $[1, 2, 1, 2, \ldots] = [\overline{1, 2}]$. What does it converge to?

$$[\overline{1,2}] = 1 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{1 + \cdots}}}},$$

so if $\alpha = [\overline{1, 2}]$ then

$$\alpha = 1 + \cfrac{1}{2 + \cfrac{1}{\alpha}}. = 1 + \cfrac{1}{\cfrac{2\alpha + 1}{\alpha}} = 1 + \frac{\alpha}{2\alpha + 1} = \frac{3\alpha + 1}{2\alpha + 1}.$$

Thus $2\alpha^2 - 2\alpha - 1 = 0$, so

$$\alpha = \frac{1 + \sqrt{3}}{2}.$$

Theorem 7.3.4. *An infinite integral continued fraction is periodic if and only if it represents a quadratic irrational.*

Proof. ($\Longrightarrow$) First suppose that

$$[a_0, a_1, \ldots, a_n, \overline{a_{n+1}, \ldots, a_{n+h}}]$$

is a periodic continued fraction. Set $\alpha = [a_{n+1}, a_{n+2}, \ldots]$. Then

$$\alpha = [a_{n+1}, \ldots, a_{n+h}, \alpha],$$

so by Proposition 7.1.3

$$\alpha = \frac{\alpha p_{n+h} + p_{n+h-1}}{\alpha q_{n+h} + q_{n+h-1}}.$$

(We use that α is the last partial convergent.) Thus α satisfies a quadratic equation. Since the a_i are all integers, the number

$$\begin{aligned} [a_0, a_1, \ldots] &= [a_0, a_1, \ldots, a_n, \alpha] \\ &= a_0 + \cfrac{1}{a_1 + \frac{1}{a_2 + \cdots + \alpha}} \end{aligned}$$

can be expressed as a polynomial in α with rational coefficients, so $[a_0, a_1, \ldots]$ also satisfies a quadratic polynomial. Finally, $\alpha \notin \mathbb{Q}$ because periodic continued fractions have infinitely many terms (the continued fraction algorithm applied to the value of an infinite integral continued fraction does not terminate).

($\Longleftarrow$) This direction was first proved by Lagrange. The proof is much more exciting than the proof of ($\Longrightarrow$)! Suppose $\alpha \in \mathbb{R}$ satisfies a quadratic equation

$$a\alpha^2 + b\alpha + c = 0$$

with $a, b, c \in \mathbb{Z}$. Let $[a_0, a_1, \ldots]$ be the continued fraction expansion of α. For each n, let

$$r_n = [a_n, a_{n+1}, \ldots],$$

so that

$$\alpha = [a_0, a_1, \ldots, a_{n-1}, r_n].$$

Our goal is to prove that the set of all r_n is finite, because then periodicity will follow easily. We have

$$\alpha = \frac{p_n}{q_n} = \frac{r_n p_{n-1} + p_{n-2}}{r_n q_{n-1} + q_{n-2}}.$$

Substituting this expression for α into the quadratic equation for α, we see that

$$A_n r_n^2 + B_n r_n + C_n = 0,$$

where

$$\begin{aligned}
A_n &= a p_{n-1}^2 + b p_{n-1} q_{n-1} + c q_{n-1}^2, \\
B_n &= 2a p_{n-1} p_{n-2} + b(p_{n-1} q_{n-2} + p_{n-2} q_{n-1}) + 2c q_{n-1} q_{n-2}, \\
C_n &= a p_{n-2}^2 + b p_{n-2} q_{n-2} + c p_{n-2}^2.
\end{aligned}$$

Note that $A_n, B_n, C_n \in \mathbb{Z}$, that $C_n = A_{n-1}$, and that

$$B^2 - 4A_n C_n = (b^2 - 4ac)(p_{n-1} q_{n-2} - q_{n-1} p_{n-2})^2 = b^2 - 4ac.$$

Recall from the proof of Theorem 7.2.9 that

$$\left| \alpha - \frac{p_{n-1}}{q_{n-1}} \right| < \frac{1}{q_n q_{n-1}}.$$

Thus

$$|\alpha q_{n-1} - p_{n-1}| < \frac{1}{q_n} < \frac{1}{q_{n-1}},$$

so

$$p_{n-1} = \alpha q_{n-1} + \frac{\delta}{q_{n-1}} \qquad \text{with } |\delta| < 1.$$

Hence

$$\begin{aligned}
A_n &= a\left(\alpha q_{n-1} + \frac{\delta}{q_{n-1}}\right)^2 + b\left(\alpha q_{n-1} + \frac{\delta}{q_{n-1}}\right) q_{n-1} + c q_{n-1}^2 \\
&= (a\alpha^2 + b\alpha + c) q_{n-1}^2 + 2a\alpha\delta + a\frac{\delta^2}{q_{n-1}^2} + b\delta \\
&= 2a\alpha\delta + a\frac{\delta^2}{q_{n-1}^2} + b\delta.
\end{aligned}$$

Thus

$$|A_n| = \left| 2a\alpha\delta + a\frac{\delta^2}{q_{n-1}^2} + b\delta \right| < 2|a\alpha| + |a| + |b|.$$

Thus there are only finitely many possibilities for the integer A_n. Also,

$$|C_n| = |A_{n-1}| \quad \text{and} \quad |B_n| = \sqrt{b^2 - 4(ac - A_n C_n)},$$

so there are only finitely many triples (A_n, B_n, C_n), and hence only finitely many possibilities for r_n as n varies. Thus there exists n and $h > 0$ such that

$$r_n = r_{n+h},$$

so

$$[a_{n+h}, a_{n+h+1}, \ldots] = [a_n, a_{n+1}, \ldots]$$

hence

$$\begin{aligned}[a_n, a_{n+1}, \ldots] &= [a_n, a_{n+1}, \ldots, a_{n+h}, \ldots] \\ &= [a_n, a_{n+1}, \ldots, a_n, a_{n+1}, \ldots] \\ &= [\overline{a_n, \ldots, a_{n+h-1}}].\end{aligned}$$

It follows that the continued fraction for α is periodic. $\square$

7.3.3 Higher Degree

Definition 7.3.5. An *algebraic number* is a root of a polynomial $f \in \mathbb{Q}[x]$.

Open Problem 7.3.6. *Give a simple description of the complete continued fraction expansion of the algebraic number $\sqrt[3]{2}$. It begins*

$$[1, 3, 1, 5, 1, 1, 4, 1, 1, 8, 1, 14, 1, 10, 2, 1, 4, 12, 2, 3, 2, 1, 3, 4, 1, 1, 2, 14, \\ 3, 12, 1, 15, 3, 1, 4, 534, 1, 1, 5, 1, 1, \ldots]$$

I don't see a pattern, and that 534 reduces my confidence that I will. One could at least try to analyze the first few terms of the continued fraction statistically (see [LT72] which suggests that $\sqrt[3]{2}$ has an "unusual" continued fraction, and [LT74] that suggests that maybe it doesn't.)

Khintchine (see [Khi63, pg. 59])

> No properties of the representing continued fractions, analogous to those which have just been proved, are known for algebraic numbers of higher degree [as of 1963]. [...] It is of interest to point out that up till the present time no continued fraction development of an algebraic number of higher degree than the second is known. It is not even known if such a development has bounded elements. Generally speaking the problems associated with the continued fraction expansion of algebraic numbers of degree higher than the second are extremely difficult and virtually unstudied.

Richard Guy (see [Guy94, pg. 260])

> Is there an algebraic number of degree greater than two whose simple continued fraction has unbounded partial quotients? Does *every* such number have unbounded partial quotients?

7.4 Applications

In this section we will learn about two applications of continued fractions. The first is a solution to the computational problem of recognizing a rational number using a computer. The second application is to the solution of "Pell's Equation": Given a positive nonsquare integer d, find *integers* x and y such that $x^2 - dy^2 = 1$.

7.4.1 Recognizing Rational Numbers

Suppose that you can compute approximations to a rational number using a computer, and desparately want to know what the rational number is. Henri Cohen gives a superb explanation in [Coh93] of how continued fraction are helpful in recognizing rational numbers.

> Consider the following apparently simple problem. Let $x \in \mathbb{R}$ be given by an approximation (for example a decimal or binary one). Decide if x is a rational number or not. Of course, this question as posed does not really make sense, since an approximation is usually itself a rational number. In practice however the question does make a lot of sense in many different contexts, and we can make it algorithmically more precise. For example, assume that one has an algorithm which allows us to compute x to as many decimal places as one likes (this is usually the case). Then, if one claims that x is (approximately) equal to a rational number p/q, this means that p/q should still be extremely close to x whatever the number of decimals asked for, p and q being fixed. This is still not completely rigorous, but it comes quite close to actual practice, so we will be content with this notion.
>
> Now how does one find p and q if x is indeed a rational number? The standard (and algorithmically excellent) answer is to compute the continued fraction expansion $[a_0, a_1, \ldots]$ of x. The number x is rational if and only if its continued fraction expansion is finite, i.e., if and only if one of the a_i is *infinite*. Since x is only given with the finite precision, x will be considered rational if x has a *very* large partial quotient a_i in its continued fraction expansion.

The following example illustrates Cohen's remarks:

Example 7.4.1. Let

$$x = 9495/3847 = 2.46815700545879906420587470756433584611385\ldots.$$

The continued fraction of the truncation 2.468157005458799064 is

$$[2, 2, 7, 2, 1, 5, 1, 1, 1, 1, 1, 1, 328210621945, 2, 1, 1, 1, \ldots]$$

We have

$$[2, 2, 7, 2, 1, 5, 1, 1, 1, 1, 1, 1] = \frac{9495}{3847}.$$

Notice that no repeat is evident in the digits of x given above, though we know that the decimal expansion of x must be eventually periodic, since all decimal expansions of fractions are eventually periodic. In fact, the length of the period of the decimal expansion of $1/3847$ is 3846 (the order of 10 modulo 3847; see Exercise 14).

7.4.2 Pell's Equation

In February of 1657, Pierre Fermat issued the following challenge:

> Given an integer $d > 1$, give a systematic way to find a positive integer y such that $dy^2 + 1$ is a perfect square.

In other words, find a solution to $x^2 - dy^2 = 1$ with $y \in \mathbb{N}$.

Note Fermat's emphasis on *integer* solutions. It is easy to find rational solutions to the equation $x^2 - dy^2 = 1$. Simply divide the relation

$$(r^2 + d)^2 - d(2r)^2 = (r^2 - d)^2$$

by $(r^2 - d)^2$ to arrive at

$$x = \frac{r^2 + d}{r^2 - d}, \qquad y = \frac{2r}{r^2 - d}.$$

Fermat said: "Solutions in fractions, which can be given at once from the merest elements of arithmetic, do not satisfy me."

The equation $x^2 - dy^2 = 1$ is called **Pell's equation**. This is because Euler (in about 1759) accidently called it "Pell's equation" and the name stuck, though Pell (1611–1685) had nothing to do with it.

If d is a perfect square, $d = n^2$, then

$$(x + ny)(x - ny) = x^2 - dy^2 = 1$$

which implies that $x + ny = x - ny = 1$, so

$$x = \frac{x + ny + x - ny}{2} = \frac{1 + 1}{2} = 1,$$

and $y = 0$ as well. We will always assume that d is not a perfect square.

7.4.3 Units in Real Quadratic Fields

From an algebraic point of view, Pell's equation is best understood in terms of units in real quadratic fields.

Let d be a nonsquare positive integer. Set

$$\mathbb{Q}(\sqrt{d}) = \{a + b\sqrt{d} : a, b \in \mathbb{Q}\} \quad \text{and} \quad \mathbb{Z}[\sqrt{d}] = \{a + b\sqrt{d} : a, b \in \mathbb{Z}\}.$$

Then $\mathbb{Q}(\sqrt{d})$ is a *real quadratic field* and $\mathbb{Z}[\sqrt{d}]$ is a ring. There is a homomorphism called norm:

$$N : \mathbb{Q}(\sqrt{d})^\times \to \mathbb{Q}^\times, \qquad N\left(a + b\sqrt{d}\right) = \left(a + b\sqrt{d}\right)\left(a - b\sqrt{d}\right) = a^2 - b^2 d.$$

Definition 7.4.2. An element $x \in R$ is a *unit* if there exists $y \in R$ such that $xy = 1$.

Proposition 7.4.3. *The units of $\mathbb{Z}[\sqrt{d}]$ are exactly the elements of norm ± 1 in $\mathbb{Z}[\sqrt{d}]$.*

Proof. First suppose $u \in \mathbb{Z}[\sqrt{d}]$ is a unit. Then

$$1 = N(1) = N(uu^{-1}) = N(u) \cdot N(u^{-1}).$$

Since $N(u), N(u^{-1}) \in \mathbb{Z}$, we have $N(u) = N(u^{-1}) = \pm 1$.

Next suppose $a + b\sqrt{d}$ has norm ± 1. Then $(a + b\sqrt{d})(a - b\sqrt{d}) = \pm 1$, so $\pm(a - b\sqrt{d})$ is an inverse of $a + b\sqrt{d}$, so $a + b\sqrt{d}$ is a unit. □

Fermat's challenge amounts to determing the group U^+ of units in $\mathbb{Z}[\sqrt{d}]$ of the form $a + b\sqrt{d}$ with $a, b \geq 0$. We will prove part of the following theorem in Section 7.4.4.

Theorem 7.4.4. *The group U^+ is an infinite cyclic group. It is generated by $p_m + q_m\sqrt{d}$, where $\frac{p_m}{q_m}$ is one of the partial convergents of the continued fraction expansion of $\sqrt{d}$. (In fact, if n is the period of the continued fraction of $\sqrt{d}$ then $m = n - 1$ when n is even and $2n - 1$ when n is odd.)*

The theorem implies that *Pell's equation always has a solution*! Warning: the smallest solution is typically shockingly large. For example, the value of x in the smallest solution to $x^2 - 1000099y^2 = 1$ has **1118 digits**. For more on how to deal with huge solutions, see Lenstra's beautiful article [Len02].

The following example illustrates how to use Theorem 7.4.4 to solve Pell's equation when $d = 61$, where the simplest solution is already quite large.

Example 7.4.5. Suppose $d = 61$. Then

$$\sqrt{d} = [7, \overline{1, 4, 3, 1, 2, 2, 1, 3, 4, 1, 14}],$$

which has odd period $n = 11$. Thus Theorem 7.4.4 asserts that U^+ is generated by

$$x = p_{21} = 1766319049$$
$$y = q_{21} = 226153980.$$

That is, we have

$$U^+ = \langle u \rangle = \langle 1766319049 + 226153980\sqrt{61} \rangle,$$

and $x = 1766319049, y = 226153980$ is a solution to $x^2 - dy^2 = 1$. All other solutions arise from u^n for some n. For example,

$$u^2 = 6239765965720528801 + 798920165762330040\sqrt{61}$$

leads to another solution.

Remark 7.4.6. Let n be an integer with $n \neq -1, 0, 1$. If the equation

$$x^2 - dy^2 = n$$

has at least one (nonzero) solution $(x_0, y_0) \in \mathbb{Z} \times \mathbb{Z}$, then it must have infinitely many. This is because if $x_0^2 - dy_0^2 = n$ and u is a generator of the cyclic group U^+, then for any integer i,

$$N(u^i(x_0 + y_0\sqrt{d})) = N(u^i) \cdot N(x_0 + y_0\sqrt{d}) = 1 \cdot n = n,$$

so

$$x_1 + y_1\sqrt{d} = u^i(x_0 + y_0\sqrt{d})$$

provides another solution to $x^2 + dy^2 = n$.

7.4.4 Some Proofs

The rest of this section is devoted to proving most of Theorem 7.4.4. We will prove that certain partial convergents to continued fractions contribute infinitely many solutions to Pell's equation. We will not prove that every solution to Pell's equation is a partial convergent, though this is true (see, e.g., [Bur89, §13.5]).

Fix a positive nonsquare integer d.

Definition 7.4.7. A quadratic irrational $\alpha = a + b\sqrt{d}$ is *reduced* if $\alpha > 1$ and if the conjugate of α, denoted by α', satisfies $-1 < \alpha' < 0$.

For example, the number $\alpha = 1 + \sqrt{2}$ is reduced.

Definition 7.4.8. A continued fraction is *purely periodic* if it is of the form $[\overline{a_0, a_1, \ldots, a_n}]$.

The continued fraction $[\overline{2}]$ of $1 + \sqrt{2}$ is purely periodic.

Lemma 7.4.9. *If α is a reduced quadratic irrational, then the continued fraction expansion of α is purely periodic. (The converse is also easily seen to be true.)*

[1] 1

Lemma 7.4.10. *The continued fraction expansion of $\sqrt{d}$ is of the form*

$$[a_0, \overline{a_1, \ldots, a_{n-1}, 2a_0}].$$

Proof. Let a_0 be the floor of $\sqrt{d}$. Then $\alpha = \sqrt{d} + a_0$ is reduced because $\alpha > 1$ and $\alpha' = -\sqrt{d} + a_0$ satisfies $-1 < \alpha' < 0$. Let $[a_0, a_1, a_2, \ldots]$ be the continued fraction expansion of $\sqrt{d}$. Then the continued fraction expansion of $\sqrt{d} + a_0$ is $[2a_0, a_1, a_2, \ldots]$. By Lemma 7.4.9, the continued fraction expansion of $\sqrt{d} + a_0$ is purely periodic, so

$$[2a_0, a_1, a_2, \ldots] = [\overline{2a_0, a_1, a_2, \ldots, a_{n-1}}],$$

where n is the period. It follows that $a_n = 2a_0$, as claimed. $\square$

[1] Give a proof of this lemma!

The following proposition shows that there are infinitely many solutions to Pell's equation that arise from continued fractions.

Proposition 7.4.11. *Let p_k/q_k be the partial convergents of the continued fraction expansion of $\sqrt{d}$, and let n be the period of the expansion of $\sqrt{d}$. Then*
$$p_{kn-1}^2 - dq_{kn-1}^2 = (-1)^{kn}$$
for $k = 1, 2, 3, \ldots$.

Proof. (This proof is taken from [Bur89, §13.5].) By Lemma 7.4.10, for $k \geq 1$, the continued fraction of $\sqrt{d}$ can be written in the form
$$\sqrt{d} = [a_0, a_1, a_2, \ldots, a_{kn-1}, r_{kn}]$$
where
$$r_{kn} = [\overline{2a_0, a_1, a_2, \ldots, a_n}] = a_0 + \sqrt{d}.$$
Because $\sqrt{d}$ is the last partial convergent of the continued fraction above, we have
$$\sqrt{d} = \frac{r_{kn}p_{kn-1} + p_{kn-2}}{r_{kn}q_{kn-1} + q_{kn-2}}.$$
Upon substituting $r_{kn} = a_0 + \sqrt{d}$ and simplifying, this reduces to
$$\sqrt{d}(a_0 a_{kn-1} + q_{kn-2} - p_{kn-1}) = a_0 p_{kn-1} + p_{kn-2} - dq_{kn-1}.$$
Because the right-hand side is rational and $\sqrt{d}$ is irrational,
$$a_0 a_{kn-1} + q_{kn-2} = p_{kn-1}, \qquad \text{and} \qquad a_0 p_{kn-1} + p_{kn-2} = dq_{kn-1}.$$
Multiplying the first of these equations by p_{kn-1} and the second by $-q_{kn-1}$, and then adding them, gives
$$p_{kn-1}^2 - dq_{kn-1}^2 = p_{kn-1}q_{kn-2} - q_{kn-1}p_{kn-2}.$$
But
$$p_{kn-1}q_{kn-2} - q_{kn-1}p_{kn-2} = (-1)^{kn-2} = (-1)^{kn},$$
which proves the proposition. □

Exercises

7.1 Compute the p_n and q_n for the continued fractions $[-3, 1, 1, 1, 1, 3]$ and $[0, 2, 4, 1, 8, 2]$. Observe that the propositions in Section 7.1.1 hold.

7.2 If $c_n = p_n/q_n$ is the nth convergent of the continued fraction $[a_0, a_1, \ldots, a_n]$ and $a_0 > 0$, show that
$$[a_n, a_{n-1}, \ldots, a_1, a_0] = \frac{p_n}{p_{n-1}}$$
and
$$[a_n, a_{n-1}, \ldots, a_2, a_1] = \frac{q_n}{q_{n-1}}.$$
(Hint: In the first case, notice that $\frac{p_n}{p_{n-1}} = a_n + \frac{p_{n-2}}{p_{n-1}} = a_n + \frac{1}{\frac{p_{n-1}}{p_{n-2}}}$.)

7.3 Evaluate each of the following infinite continued fractions:

(a) $[\overline{2,3}]$

(b) $[2,\overline{1,2,1}]$

(c) $[0,\overline{1,2,3}]$

7.4 Determine the infinite continued fraction of each of the following numbers:

(a) $\sqrt{5}$

(b) $\dfrac{1+\sqrt{13}}{2}$

(c) $\dfrac{5+\sqrt{37}}{4}$

7.5 (a) For any positive integer n, prove that $\sqrt{n^2+1} = [n,\overline{2n}]$.

(b) Find a convergent to $\sqrt{5}$ that approximates $\sqrt{5}$ to within four decimal places.

7.6 A theorem of Hurwitz (1891) asserts that for any irrational number x, there exists infinitely many rational numbers a/b such that

$$\left|x - \frac{a}{b}\right| < \frac{1}{\sqrt{5}b^2}.$$

Take $x = e$, and obtain four rational numbers that satisfy this inequality.

7.7 The continued fraction expansion of e is

$$[2, 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, 1, 1, \ldots].$$

It is a theorem that the obvious pattern continues indefinitely. Do you think that the continued fraction expansion of e^2 also exhibits a nice pattern? If so, what do you think it is?

7.8 (a) Show that there are infinitely many even integers n with the property that both $n+1$ and $\frac{n}{2}+1$ are perfect squares.

(b) Exhibit two such integers that are greater than 389.

7.9 A primitive Pythagorean triple is a triple x, y, z of integers such that $x^2 + y^2 = z^2$. Prove that there exists infinitely many primitive Pythagorean triples x, y, z in which x and y are consecutive integers.

7.10 Find two distinct continued fractions $a_0, a_1, a_2, \ldots$ and $b_0, b_1, b_2, \ldots$ such that

$$[a_0, a_1, a_2, \ldots] = [b_0, b_1, b_2, \ldots].$$

(Note that necessarily the a_i and b_i won't all be integers.)

7.11 (a) Find the continued fraction expansion of $(1+2\sqrt{3})/4$. Prove that your answer is correct.

(b) Evaluate the infinite continued fraction $[0, \overline{1, 3}]$

7.12 Let $a_0 \in \mathbb{R}$ and $a_1, \ldots, a_n$ and b be positive real numbers. Prove that

$$[a_0, a_1, \ldots, a_n + b] < [a_0, a_1, \ldots, a_n]$$

if and only if n is odd.

7.13 Let $s(n) = 1 + 2 + \cdots + n = \frac{n(n+1)}{2}$ be the sum of the first n positive integers. Prove that there are infinitely many n such that $s(n)$ is a perfect square. (Hint: Find a relationship between such n and solutions to a certain Pell's equation.)

7.14 (a) Prove that for every positive integer r,

$$\frac{1}{1 - 10^r} = \sum_{n \geq 1} 10^{-rn}.$$

(b) Let d be an integer that is coprime to 10. Prove that the decimal expansion of $\frac{1}{d}$ has period equal to the order of 10 modulo d.

7.15 Let α be a real number, and let p_k/q_k denote the partial convergents of the integral continued fraction for α.

(a) Prove that for every $k \geq 0$,

$$\left| \alpha - \frac{p_k}{q_k} \right| < q_k^2.$$

(b) Let the decimal expansion of α be

$$\alpha = b + \frac{b_1}{10} + \frac{b_2}{10^2} + \frac{b_3}{10^3} + \frac{b_4}{10^4} + \cdots,$$

where $0 \leq b_n \leq 9$ for all n. Suppose that for some convergent p_k/q_k we have $q_k = 100$. Prove that either $b_3 = b_4 = 0$ or $b_3 = b_4 = 9$. (This problem is inspired by [Sta78, pg. 210].)

8
Adic Numbers

8.1 The N-Adic Numbers

Convince yourself that the following lemma is true.

Lemma 8.1.1. *Let $N > 1$ be an integer. Then for any positive rational number α there exists unique $e \in \mathbb{Z}$ and positive integers a, b such that $\alpha = N^e \cdot \frac{a}{b}$ with $N \nmid a$, $\gcd(a,b) = 1$, and $\gcd(N,b) = 1$.*

Definition 8.1.2 (N-adic valuation). Let N be a positive integer. For any positive $\alpha \in \mathbb{Q}$, the *N-adic valuation* of α is e, where e is as in Lemma 8.1.1. If α is negative, the N-adic valuation of α is the valuation of $-\alpha$. The N-adic valuation of 0 is ∞.

We denote the N-adic valuation of α by $v_N(\alpha)$.

Definition 8.1.3 (N-adic metric). For $x, y \in \mathbb{Q}$ the *N-adic distance* between x and y is

$$d_N(x,y) = N^{-v_N(x-y)}.$$

We let $d_N(x,x) = 0$, since $v_N(x-x) = v_N(0) = \infty$.

For example, $x, y \in \mathbb{Z}$ are close in the N-adic metric if their difference is divisible by a large power of N. E.g., if $N = 10$ then 93427 and 13427 are close because their difference is 80000, which is divisible by a large power of 10.

Definition 8.1.4 (Metric). A *metric* on a set X is a map

$$d : X \times X \to \mathbb{R}$$

such that for all $x, y, z \in X$,

1. $d(x,y) \geq 0$ and $d(x,y) = 0$ if and only if $x = y$,

2. $d(x, y) = d(y, x)$, and
3. $d(x, z) \leq d(x, y) + d(y, z)$.

We recall from a basic analysis course the following facts about completion with respect to a metric. A *Cauchy sequence* is a sequence (x_n) in X such that for all $\varepsilon > 0$ there exists M such that for all $n, m > M$ we have $d(x_n, x_m) < \varepsilon$. The *completion* of X is the set of Cauchy sequences (x_n) in X modulo the equivalence relation in which two Cauchy sequences (x_n) and (y_n) are equivalent if $\lim_{n\to\infty} d(x_n, y_n) = 0$. A metric space is *complete* if every Cauchy sequence converges, and one can show that the completion of X with respect to a metric is complete.

For example, $d(x, y) = |x - y|$ defines a metric on $\mathbb{Q}$. The completion of $\mathbb{Q}$ with respect to this metric is the field $\mathbb{R}$ of real numbers. In certain parts of number theory the N-adic numbers, which we introduce shortly, are just as important as $\mathbb{R}$.

Proposition 8.1.5. *The distance d_N on $\mathbb{Q}$ defined above is a metric. Moreover, for all $x, y, z \in \mathbb{Q}$ we have*

$$d(x, z) \leq \max(d(x, y), d(y, z)).$$

Proof. The first two properties of Definition 8.1.4 are immediate. For the third, we first prove that if $\alpha, \beta \in \mathbb{Q}$ then

$$v_N(\alpha + \beta) \geq \min(v_N(\alpha), v_N(\beta)).$$

Assume, without loss, that $v_N(\alpha) \leq v_N(\beta)$ and that both α and β are nonzero. Using Lemma 8.1.1 write $\alpha = N^e(a/b)$ and $\beta = N^f(c/d)$ with a or c possibly negative. Then

$$\alpha + \beta = N^e\left(\frac{a}{b} + N^{f-e}\frac{c}{d}\right) = N^e\left(\frac{ad + bcN^{f-e}}{bd}\right).$$

Since $\gcd(N, bd) = 1$ it follows that $v_N(\alpha+\beta) \geq e$. Now suppose $x, y, z \in \mathbb{Q}$. Then

$$x - z = (x - y) + (y - z),$$

so

$$v_N(x - z) \geq \min(v_N(x - y), v_N(y - z)),$$

hence $d_N(x, z) \leq \max(d_N(x, y), d_N(y, z))$. □

We can finally define the N-adic numbers.

Definition 8.1.6 (The N-adic Numbers). The set of *N-adic numbers*, denoted $\mathbb{Q}_N$, is the completion of $\mathbb{Q}$ with respect to the metric d_N.

The set $\mathbb{Q}_N$ is a ring, but it need not be a field as you will show in Exercises 4 and 5. Also, $\mathbb{Q}_N$ has a bizarre topology, as we will see in Section 8.4.

8.2 The 10-Adic Numbers

It's a familiar fact that every real number can be written in the form

$$d_n \dots d_1 d_0 . d_{-1} d_{-2} \dots = d_n 10^n + \dots + d_1 10 + d_0 + d_{-1} 10^{-1} + d_{-2} 10^{-2} + \dots$$

where each digit d_i is between 0 and 9, and the sequence can continue indefinitely to the right.

The 10-adic numbers also have decimal expansions, but everything is backwards! To get a feeling for why this might be the case, we consider Euler's nonsensical series

$$\sum_{n=1}^{\infty} (-1)^{n+1} n! = 1! - 2! + 3! - 4! + 5! - 6! + \cdots .$$

You will prove in Exercise 2 that this series converges in $\mathbb{Q}_{10}$ to some element $\alpha \in \mathbb{Q}_{10}$.

What is α? How can we write it down? First note that for all $M \geq 5$, the terms of the sum are divisible by 10, so the difference between α and $1! - 2! + 3! - 4!$ is divisible by 10. Thus we can compute α modulo 10 by computing $1! - 2! + 3! - 4!$ modulo 10. Likewise, we can compute α modulo 100 by compute $1! - 2! + \cdots + 9! - 10!$, etc. We obtain the following table:

α	mod 10^r
1	mod 10
81	mod 10^2
981	mod 10^3
2981	mod 10^4
22981	mod 10^5
422981	mod 10^6

Continuing we see that

$$1! - 2! + 3! - 4! + \cdots = \dots 637838364422981 \qquad \text{in } \mathbb{Q}_{10} \ !$$

Here's another example. Reducing $1/7$ modulo larger and larger powers of 10 we see that

$$\frac{1}{7} = \dots 857142857143 \qquad \text{in } \mathbb{Q}_{10}.$$

Here's another example, but with a decimal point.

$$\frac{1}{70} = \frac{1}{10} \cdot \frac{1}{7} = \dots 85714285714.3$$

We have

$$\frac{1}{3} + \frac{1}{7} = \dots 66667 + \dots 57143 = \frac{10}{21} = \dots 23810,$$

which illustrates that addition with carrying works as usual.

8.2.1 FLT in $\mathbb{Q}_{10}$

An amusing observation, which people used to endlessly argue about on USENET back in the 1990s, is that Fermat's last theorem is false in $\mathbb{Q}_{10}$. For example, $x^3 + y^3 = z^3$ has a nontrivial solution, namely $x = 1$, $y = 2$, and $z = \ldots 60569$. Here z is a cube root of 9 in $\mathbb{Q}_{10}$. Note that it takes some work to prove that there is a cube root of 9 in $\mathbb{Q}_{10}$ (see Exercise 3).

8.3 The Field of p-Adic Numbers

The ring $\mathbb{Q}_{10}$ of 10-adic numbers is isomorphic to $\mathbb{Q}_2 \times \mathbb{Q}_5$ (see Exercise 5), so it is not a field. For example, the element $\ldots 8212890625$ corresponding to $(1, 0)$ under this isomorphism has no inverse. (To compute n digits of $(1, 0)$ use the Chinese remainder theorem to find a number that is 1 modulo 2^n and 0 modulo 5^n.)

If p is prime then $\mathbb{Q}_p$ is a field (see Exercise 4). Since $p \neq 10$ it is a little more complicated to write p-adic numbers down. People typically write p-adic numbers in the form

$$\frac{a_{-d}}{p^d} + \cdots + \frac{a_{-1}}{p} + a_0 + a_1 p + a_2 p^2 + a_3 p^3 + \cdots$$

where $0 \leq a_i < p$ for each i.

8.4 The Topology of $\mathbb{Q}_N$ (is Weird)

Definition 8.4.1 (Connected). Let X be a topological space. A subset S of X is *disconnected* if there exist open subsets $U_1, U_2 \subset X$ with $U_1 \cap U_2 \cap S = \emptyset$ and $S = (S \cap U_1) \cup (S \cap U_2)$ with $S \cap U_1$ and $S \cap U_2$ nonempty. If S is not disconnected it is *connected*.

The topology on $\mathbb{Q}_N$ is induced by d_N, so every open set is a union of open balls

$$B(x, r) = \{y \in \mathbb{Q}_N : d_N(x, y) < r\}.$$

Recall Proposition 8.1.5, which asserts that for all x, y, z,

$$d(x, z) \leq \max(d(x, y), d(y, z)).$$

This translates into the following shocking and bizarre lemma:

Lemma 8.4.2. *Suppose* $x \in \mathbb{Q}_N$ *and* $r > 0$. *If* $y \in \mathbb{Q}_N$ *and* $d_N(x, y) \geq r$, *then* $B(x, r) \cap B(y, r) = \emptyset$.

Proof. Suppose $z \in B(x, r)$ and $z \in B(y, r)$. Then

$$r \leq d_N(x, y) \leq \max(d_N(x, z), d_N(z, y)) < r,$$

a contradiction. $\square$

You should immediately draw a picture that illustrates Lemma 8.4.2.

Lemma 8.4.3. *The open ball $B(x,r)$ is also closed.*

Proof. Suppose $y \notin B(x,r)$. Then $r < d(x,y)$ so

$$B(y, d(x,y)) \cap B(x,r) \subset B(y, d(x,y)) \cap B(x, d(x,y)) = \emptyset.$$

Thus the complement of $B(x,r)$ is a union of open balls. □

The lemmas imply that $\mathbb{Q}_N$ is *totally disconnected*, in the following sense.

Proposition 8.4.4. *The only connected subsets of $\mathbb{Q}_N$ are the singleton sets $\{x\}$ for $x \in \mathbb{Q}_N$ and the empty set.*

Proof. Suppose $S \subset \mathbb{Q}_N$ is a nonempty connected set and x, y are distinct elements of S. Let $r = d_N(x,y) > 0$. Let $U_1 = B(x,r)$ and U_2 be the complement of U_1, which is open by Lemma 8.4.3. Then U_1 and U_2 satisfies the conditions of Definition 8.4.1, so S is not connected, a contradiction. □

8.5 The Local-to-Global Principle of Hasse and Mikowski

Section 8.4 might have convinced you that $\mathbb{Q}_N$ is a bizarre pathology. In fact, $\mathbb{Q}_N$ is omnipresent in number theory, as the following two fundamental examples illustrate.

A “nontrivial” solution to a homogeneous equation is a solution where not all indeterminates are 0.

Theorem 8.5.1 (Hasse-Minkowski). *The quadratic equation*

$$a_1x_1^2 + a_2x_2^2 + \cdots + a_nx_n^2 = 0 \tag{8.1}$$

with $a_i \in \mathbb{Q}^\times$ has a nontrivial solution with $x_1, \ldots, x_n$ in $\mathbb{Q}$ if and only if (8.1) has a solution in $\mathbb{R}$ and in $\mathbb{Q}_p$ for all primes p.

This theorem is very useful in practice because the p-adic condition turns out to be easy to check. For more details, including a complete proof, see IV.3.2 of Serre's *A Course In Arithmetic.*

The analogue of Theorem 8.5.1 for cubic equations is false. For example, Selmer proved that the cubic

$$3x^3 + 4y^3 + 5z^3 = 0$$

has a nontrivial solution in $\mathbb{R}$ and in $\mathbb{Q}_p$ for all primes p but has no solutions in $\mathbb{Q}$.

Open Problem 8.5.2. *Give an algorithm that decides whether or not a cubic $ax^3 + by^3 + cz^3 = 0$ has a nontrivial solution in $\mathbb{Q}$.*

This open problem is closely related to the Birch and Swinnerton-Dyer Conjecture for elliptic curves. The truth of the conjecture would follow if we knew that “Shafarevich-Tate Groups” of elliptic curves are finite.

EXERCISES

8.1 Compute the first 5 digits of the 10-adic expansions of the following rational numbers:

$$\frac{13}{2}, \quad \frac{1}{389}, \quad \frac{17}{19}, \qquad \text{the 4 square roots of 41.}$$

8.2 Let $N > 1$ be an integer. Prove that the series

$$\sum_{n=1}^{\infty}(-1)^{n+1}n! = 1! - 2! + 3! - 4! + 5! - 6! + \cdots.$$

converges in $\mathbb{Q}_N$.

8.3 Prove that -9 has a cube root in $\mathbb{Q}_{10}$ using the following strategy (this is a special case of "Hensel's Lemma").

(a) Show that there is $\alpha \in \mathbb{Z}$ such that $\alpha^3 \equiv 9 \pmod{10^3}$.

(b) Suppose $n \geq 3$. Use induction to show that if $\alpha_1 \in \mathbb{Z}$ and $\alpha^3 \equiv 9 \pmod{10^n}$, then there exists $\alpha_2 \in \mathbb{Z}$ such that $\alpha_2^3 \equiv 9 \pmod{10^{n+1}}$. (Hint: Show that there is an integer b such that $(\alpha_1 + b10^n)^3 \equiv 9 \pmod{10^{n+1}}$.)

(c) Conclude that 9 has a cube root in $\mathbb{Q}_{10}$.

8.4 Let $N > 1$ be an integer.

(a) Prove that $\mathbb{Q}_N$ is equipped with a natural ring structure.

(b) If N is prime, prove that $\mathbb{Q}_N$ is a field.

8.5 (a) Let p and q be distinct primes. Prove that $\mathbb{Q}_{pq} \cong \mathbb{Q}_p \times \mathbb{Q}_q$.

(b) Is $\mathbb{Q}_{p^2}$ isomorphic to either of $\mathbb{Q}_p \times \mathbb{Q}_p$ or $\mathbb{Q}_p$?

9

Binary Quadratic Forms and Ideal Class Groups

In this chapter, we will study binary quadratic forms.[1] 1

A simple example of a binary quadratic form is

$$f(x,y) = x^2 + y^2.$$

For which integers n do there exist integer x and y such that $x^2 + y^2 = n$? We will answer this question in Section 9.1.

[2] 2

9.1 Sums of Two Squares

Theorem 9.1.1. *A number n is a sum of two squares if and only if all prime factors of n of the form $4m+3$ have even exponent in the prime factorization of n.*

In this section we give two very different proofs of Theorem 9.1.1. The first is very arithmetic and builds on results about continued fractions from Chapter 7. The second more algebraic proof uses quadratic reciprocity from Chapter 6 to understand splitting of ideals in the ring $\mathbb{Z}[i]$ of Gaussian integers.

Before tackling the proofs, we consider a few examples. Notice that $5 = 1^2 + 2^2$ is a sum of two squares, but 7 is not a sum of two squares, because the congruence $x^2 + y^2 \equiv 7 \pmod{8}$ has no solution. Since 2001 is divisible by 3 (because $2 + 1$), but not by 9 (since $2 + 1$ is not), Theorem 9.1.1 implies that 2001 is not a sum of two squares. The theorem implies that

[1] Put a complete chapter outline here, with philosophy.

[2] Say that this chapter benefited immensely from Cohn's "Advanced Number Theory".

$2 \cdot 3^4 \cdot 5 \cdot 7^2 \cdot 13$ is a sum of two squares, but that $21 = 3 \cdot 7$ is not a sum of two squares even though $21 \equiv 1 \pmod 4$.

Remark 9.1.2. More generally, every natural number is a sum of four integer squares. A natural number is a sum of three squares if and only if it is not a power of 4 times a number that is congruent to 7 modulo 8. For example, 7 is not a sum of three squares, as one can easily see by considering $x^2 + y^2 + z^2 \equiv 7 \pmod 8$. See[3] for proofs. 3

Definition 9.1.3 (Primitive). A representation $n = x^2 + y^2$ is *primitive* if x and y are coprime.

Lemma 9.1.4. *If n is divisible by a prime p of the form $4m + 3$, then n has no primitive representations.*

Proof. Suppose $p = 4m + 3$ divides n. If n has a primitive representation, $n = x^2 + y^2$, then

$$p \mid x^2 + y^2 \quad \text{and} \quad \gcd(x, y) = 1,$$

so $p \nmid x$ and $p \nmid y$. Since $\mathbb{Z}/p$ is a field we divide by y^2 in the equation $x^2 + y^2 \equiv 0 \pmod p$ to see that $(x/y)^2 \equiv -1 \pmod p$. Thus the quadratic residue symbol $\left(\frac{-1}{p}\right)$ equals $+1$. However, by Proposition 6.2.1,

$$\left(\frac{-1}{p}\right) = (-1)^{(p-1)/2} = (-1)^{(4m+3-1)/2} = (-1)^{2m+1} = -1,$$

a contradiction. Thus no prime of the form $4m + 3$ divides n. $\square$

Proof of Theorem 9.1.1 ($\Longrightarrow$). Suppose that p is of the form $4m + 3$, that $p^r \parallel n$ (i.e., $p^r \mid n$ but $p^{r+1} \nmid n$) with r odd, and that $n = x^2 + y^2$. Letting $d = \gcd(x, y)$, we have

$$x = dx', \quad y = dy', \quad n = d^2 n'$$

with $\gcd(x', y') = 1$ and

$$(x')^2 + (y')^2 = n'.$$

Because r is odd, $p \mid n'$, so Lemma 9.1.4 implies that $\gcd(x', y') > 1$, a contradiction. $\square$

To prepare for our two proofs of the ($\Longleftarrow$) direction of Theorem 9.1.1, we reduce the problem to the case when n is prime. Write $n = n_1^2 n_2$ where n_2 has no prime factors of the form $4m + 3$. It suffices to show that n_2 is a sum of two squares. Also note that

$$(x_1^2 + y_1^2)(x_2^2 + y_2^2) = (x_1x_2 - y_1y_2)^2 + (x_1y_2 + x_2y_1)^2,$$

so a product of two numbers that are sums of two squares is also a sum of two squares. (This algebraic identity is the assertion that the norm $N : \mathbb{Q}(i)^\times \to \mathbb{Q}^\times$ sending $x + iy$ to $(x + iy)(x - iy) = x^2 + y^2$ is a group homomorphism.) Also, $2 = 1^2 + 1^2$ is a sum of two squares.

It thus suffices to show that if $p = 4m + 1$, then p is a sum of two squares.

[3] Choose a nice reference. Hardy-Wright?

9.1.1 Arithmetic Proof of Theorem 9.1.1

Lemma 9.1.5. *If $x \in \mathbb{R}$ and $n \in \mathrm{N}$, then there is a fraction $\dfrac{a}{b}$ in lowest terms such that $0 < b \le n$ and*

$$\left| x - \frac{a}{b} \right| \le \frac{1}{b(n+1)}.$$

Proof. Consider the continued fraction expansion $[a_0, a_1, \ldots]$ of x. By Corollary 7.2.10, for each m

$$\left| x - \frac{p_m}{q_m} \right| < \frac{1}{q_m \cdot q_{m+1}}.$$

Since $q_{m+1} \ge q_m + 1$ and $q_0 = 1$, either there exists an m such that $q_m \le n < q_{m+1}$, or the continued fraction expansion of x is finite and n is larger than the denominator of the rational number x, in which case we take $\frac{a}{b} = x$ and are done. In the first case,

$$\left| x - \frac{p_m}{q_m} \right| < \frac{1}{q_m \cdot q_{m+1}} \le \frac{1}{q_m \cdot (n+1)},$$

so $\dfrac{a}{b} = \dfrac{p_m}{q_m}$ satisfies the conclusion of the lemma. □

Proof of Theorem 9.1.1 ($\Longleftarrow$). Suppose $p = 4m + 1$ is a prime. Since

$$(-1)^{(p-1)/2} = (-1)^{(4m+1-1)/2} = +1,$$

Proposition 6.2.1 implies that -1 is a square modulo p; i.e., there exists $r \in \mathbb{Z}$ such that $r^2 \equiv -1 \pmod{p}$. Lemma 9.1.5, with $n = \lfloor \sqrt{p} \rfloor$ and $x = -\frac{r}{p}$, implies that there are integers a, b such that $0 < b < \sqrt{p}$ and

$$\left| -\frac{r}{p} - \frac{a}{b} \right| \le \frac{1}{b(n+1)} < \frac{1}{b\sqrt{p}}.$$

Let $c = rb + pa$; then

$$|c| < \frac{pb}{b\sqrt{p}} = \frac{p}{\sqrt{p}} = \sqrt{p}$$

and

$$0 < b^2 + c^2 < 2p.$$

But $c \equiv rb \pmod{p}$, so

$$b^2 + c^2 \equiv b^2 + r^2 b^2 \equiv b^2(1 + r^2) \equiv 0 \pmod{p}.$$

Thus $b^2 + c^2 = p$. □

9.1.2 Algebraic Proof of Theorem 9.1.1

Let p be a prime that is congruent to 1 modulo 4. In this section we show that p is a sum of two squares by factoring the ideal generated by p in the Gaussian integers as a product of principal ideals.

The Gaussian integers are

$$R = \mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\} \subset \mathbb{C},$$

where $i^2 = -1$. The *ideal* generated by $x_1, \ldots, x_n \in R$ is the set (in fact, R-module)

$$(x_1, \ldots, x_n) = Rx_1 + \cdots + Rx_n \subset R$$

of R-linear combinations of the x_i.

Lemma 9.1.6. *There is an integer r such that*

$$(p) = (i - r, p)(i + r, p),$$

where the equality is an equality of ideals in R.

Proof. Because $p \equiv 1 \pmod 4$, we have $(-1)^{(p-1)/2} = 1$, so by Proposition 6.2.1, there is an $r \in \mathbb{Z}$ such that $r^2 \equiv -1 \pmod p$. The ideal product $(i - r, p)(i + r, p)$ is, by definition, the ideal generated by all products of elements in $(i - r, p)$ with elements in $(i + r, p)$. In particular, it contains p^2, $1 + r^2 = -(i - r)(i + r)$, and $-2pr = p(i - r) - p(i + r)$. Since p is odd and divides $1 + r^2$, the greatest common divisor of p^2, $1 + r^2$, and $-2pr$ is p, so $(p) \subset (i - r, p)(i + r, p)$. Since $(i - r)(i + r) = -1 - r^2$ is a multiple of p we see that every element of $(i - r, p)(i + r, p)$ is a multiple of p, which completes the proof. $\square$

The lemma is not quite enough to conclude that p is of the form $a^2 + b^2$. For that, we show that $(i - r, p)$ is generated by a single element, i.e., it is *principal*. The following proposition asserts that every ideal of R is principal by observing that an analogue of the division algorithm holds in R.

Proposition 9.1.7. *The ring R is a principal ideal domain (because it is a Euclidean domain).*

Proof. First we show that R is a Euclidean domain, i.e., there is a function $\lambda : R \to \mathbb{Z}_{\geq 0} = \mathbf{N} \cup \{0\}$ such that for all $x, y \in R$ with $x \neq 0$, there exist $q, r \in R$ such that $y = xq + r$ with $\lambda(r) < \lambda(x)$. To see this, let

$$\lambda(a + bi) = \mathrm{N}(a + bi) = a^2 + b^2$$

be the norm. Then if $x = a + bi \neq 0$ and $y = c + di$, we have

$$\frac{y}{x} = \frac{c + di}{a + bi} = \frac{ac + bd}{\mathrm{N}(x)} + \frac{ad - bc}{\mathrm{N}(x)} i.$$

Let e and f be the integers that are closest to the real and imaginary parts of y/x, respectively. Let $q = e + if$ and $r = y - xq$. Then

$$\mathrm{N}(r) = \mathrm{N}(y/x - q)\,\mathrm{N}(x) \leq \frac{1}{2} \cdot \mathrm{N}(x).$$

It is now easy to deduce that R is a principal ideal domain. Suppose $I \subset R$ is any nonzero ideal. Let x be an element of I with $\mathrm{N}(x)$ minimal. If $y \in I$ then $y = qx + r$ with $\mathrm{N}(r) < \mathrm{N}(x)$. Since $r = qx - y \in I$, it follows that $\mathrm{N}(r) = 0$, so $r = 0$ and $y \in (x)$. Thus $I = (x)$. □

Recall that

$$(p) = (i - r, p)(i + r, p).$$

By Proposition 9.1.7 the ideals on the right side are principal, so there exists $a + bi \in R$ such that

$$(p) = (a + bi)(a - bi).$$

Since $(a + bi)(a - bi) = (a^2 + b^2)$, it follows that $p = (a^2 + b^2)u$ for some unit u. The units of R are $\pm 1, \pm i$, so since p and $a^2 + b^2$ are positive real numbers, the only possibility is that $u = 1$. Thus $p = a^2 + b^2$, which completes our algebraic proof of Theorem 9.1.1.

This proof is longer than the proof in Section 9.1.1, but every step involves learning about the structure of interesting basic number-theoretic objects. Moreover, the underlying idea of the proof is clearer and suggests generalizations to the problem of representation of primes by more general expressions.

9.2 Binary Quadratic Forms

9.2.1 Introduction

A *binary quadratic form* is a homogeneous polynomial

$$f = ax^2 + bxy + cy^2 \in \mathbb{Z}[x, y].$$

We say that n is *represented* by f if there are integers $x, y \in \mathbb{Z}$ such that $f(x, y) = n$. The representability problem will partially motivate our interest in quadratic forms.

Problem 9.2.1. Given a binary quadratic form f, give a good way to determine whether or not any given integer n is represented by f.

We gave a simple solution to this problem in Section 9.1 in the case when $f = x^2 + y^2$. The set of sums of two squares is the set of integers n such that any prime divisor p of n of the form $4m + 3$ exactly divides n to an even power (along with 0).

9.2.2 Equivalence

For simplicity below we will sometimes write $f\begin{pmatrix} x \\ y \end{pmatrix}$ for $f(x, y)$.

Definition 9.2.2. The modular group $\mathrm{SL}_2(\mathbb{Z})$ is the group of all 2×2 integer matrices with determinant $+1$.

If $\gamma = \left(\begin{smallmatrix} p & q \\ r & s \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$ and $f(x, y) = ax^2 + bxy + cy^2$ is a quadratic form, let

$$f|_\gamma(x, y) = f(px + qy, rx + sy) = f\left(\begin{pmatrix} p & q \\ r & s \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}\right).$$

Proposition 9.2.3. *The above formula defines a right action of the group* $\mathrm{SL}_2(\mathbb{Z})$ *on the set of binary quadratic forms, in the sense that*

$$f|_{\gamma\delta} = (f|_\gamma)|_\delta,$$

for any $\gamma, \delta \in \mathrm{SL}_2(\mathbb{Z})$.

Proof. Suppose $\gamma, \delta \in \mathrm{SL}_2(\mathbb{Z})$. Then

$$f|_{\gamma\delta}\begin{pmatrix} x \\ y \end{pmatrix} = f\left(\gamma\delta\begin{pmatrix} x \\ y \end{pmatrix}\right) = f|_\gamma\left(\delta\begin{pmatrix} x \\ y \end{pmatrix}\right) = (f|_\gamma)|_\delta\begin{pmatrix} x \\ y \end{pmatrix}.$$

□

Proposition 9.2.4. *Let* $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ *and let* f *be a binary quadratic form. The set of integers represented by* f *is exactly the same as the set of integers represented by* $f|_\gamma$. *(The converse is not true; see Example 9.3.4.)*

Proof. If $f(x_0, y_0) = n$ then since $\gamma^{-1} \in \mathrm{SL}_2(\mathbb{Z})$, we have $\gamma^{-1}\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \in \mathbb{Z}^2$, so

$$f|_\gamma\left(\gamma^{-1}\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}\right) = f\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = n.$$

Thus every integer represented by f is also represented by $f|_\gamma$. Conversely, if $f|_\gamma \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = n$, then $f\left(\gamma \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}\right) = n$, so f represents n. $\square$

Define an equivalence relation $\sim$ on the set of all binary quadratic forms by $f \sim f'$ if there exists $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ such that $f|_\gamma = f'$.

For simplicity, we will sometimes denote the quadratic form $ax^2 + bxy + cy^2$ by (a,b,c). Then, for example, since $\gamma = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$, we see that $(a,b,c) \sim (c,-b,a)$, since if $f(x,y) = ax^2 + bxy + cy^2$, then $f(-y,x) = ay^2 - bxy + cx^2$.

Example 9.2.5. Consider the binary quadratic form

$$f(x,y) = 458x^2 + 214xy + 25y^2.$$

Solving the representation problem for f might, at first glance, look hopeless. We find $f(x,y)$ for a few values of x and y:

$$\begin{aligned} f(-1,-1) &= 17 \cdot 41 \\ f(-1,0) &= 2 \cdot 229 \\ f(0,-1) &= 5^2 \\ f(1,1) &= 269 \\ f(-1,2) &= 2 \cdot 5 \cdot 13 \\ f(-1,3) &= 41 \end{aligned}$$

Each number is a sum of two squares! Letting $\gamma = \left(\begin{smallmatrix} 4 & -3 \\ -17 & 13 \end{smallmatrix}\right)$, we have

$$\begin{aligned} f|_\gamma &= 458(4x-3y)^2 + 214(4x-3y)(-17x+13y) + 25(-17x+13y)^2 \\ &= \cdots = x^2 + y^2!! \end{aligned}$$

Thus by Proposition 9.2.4, f represents an integer n if and only if n is a sum of two squares.

9.2.3 Discriminants

Definition 9.2.6. The *discriminant* of $f(x,y) = ax^2+bxy+cy^2$ is b^2-4ac.

Example 9.2.7. Notice that $\mathrm{disc}(x^2+y^2) = -4$ and

$$\mathrm{disc}(458, 214, 25) = 214^2 - 4 \cdot 25 \cdot 458 = -4.$$

That the discriminants are the same indicates that $(1,0,1)$ and $(458,214,25)$ are closely related.

Proposition 9.2.8. *If $f \sim f'$, then* $\mathrm{disc}(f) = \mathrm{disc}(f')$.

Proof. By elementary algebra, one sees that if $\gamma \in \mathrm{SL}_2(\mathbb{Z})$, then

$$\mathrm{disc}(f|_\gamma) = \mathrm{disc}(f) \cdot \det(\gamma)^2 = \mathrm{disc}(f).$$

Since $f' = f|_\gamma$ for some $\gamma \in \mathrm{SL}_2(\mathbb{Z})$, the proposition follows. $\square$

The converse of the proposition is false. Forms with the same discriminant need not be equivalent. For example, the forms $(1,0,6)$ and $(2,0,3)$ have discriminant -24, but are not equivalent. To see this, observe that $(1,0,6)$ represents 1, but $2x^2+3y^2$ can not represent 1.

Proposition 9.2.9. *The set of all discriminants of forms is exactly the set of integers d such that $d \equiv 0$ or $1 \pmod 4$.*

Proof. First note that b^2-4ac is a square modulo 4, so it must equal 0 or 1 modulo 4. Next suppose d is an integer such that $d \equiv 0$ or $1 \pmod 4$. If we set

$$c = \begin{cases} -d/4, & \text{if } d \equiv 0 \pmod 4 \\ -(d-1)/4 & \text{if } d \equiv 1 \pmod 4, \end{cases}$$

then $\operatorname{disc}(1,0,c) = d$ in the first case and $\operatorname{disc}(1,1,c) = d$ in the second. □

Definition 9.2.10. The form $(1,0,-d/4)$ or $(1,1,-(d-1)/4)$ of discriminant d that appears in the proof of the previous proposition is called the *principal form* of discriminant d.

d	principal form	
-4	(1,0,1)	x^2+y^2
5	(1,1,-1)	x^2+xy-y^2
-7	(1,1,2)	$x^2+xy+2y^2$
8	(1,0,-2)	x^2-2y^2
-23	(1,1,6)	$x^2+xy+6y^2$
389	(1,1,-97)	$x^2+xy-97y^2$

9.2.4 Definite and Indefinite Forms

Definition 9.2.11. A quadratic form with negative discriminant is called *definite*. A form with positive discriminant is called *indefinite*.

This definition is motivated by the fact that the nonzero integers represented by a definite form are all either positive or negative. To see this, let (a,b,c) be a quadratic form, multiply by $4a$ and complete the square:

$$\begin{aligned} 4a(ax^2+bxy+cy^2) &= 4a^2x^2+4abxy+4acy^2 \\ &= (2ax+by)^2+(4ac-b^2)y^2 \end{aligned}$$

If $\operatorname{disc}(a,b,c) < 0$ then $4ac-b^2 = -\operatorname{disc}(a,b,c) > 0$, so the nonzero values taken on by $ax^2+bxy+cy^2$ are only positive or only negative, depending on the sign of a. On the other hand, if $\operatorname{disc}(a,b,c) > 0$, then $(2ax+by)^2+(4ac-b^2)y^2$ takes both positive and negative values, so (a,b,c) does also.

9.2.5 Rings of Integers in Quadratic Fields

We have seen quadratic number fields several times before. We now make the theory more precise, in order to see how the arithmetic of quadratic number fields is closely linked to the theory of quadratic forms.

Let $D \neq 0, 1$ be a square-free integer. The quadratic field obtained by adjoining $\sqrt{D}$ to $\mathbb{Q}$ is

$$K = \mathbb{Q}(\sqrt{D}) = \{a + b\sqrt{D} \mid a, b \in \mathbb{Q}\}.$$

Definition 9.2.12 (Integral). An element $x \in K$ is *integral over* $\mathbb{Z}$ if it is the root of a quadratic polynomial of the form $x^2 + ax + b$ with $a, b \in \mathbb{Z}$.

The integral elements of K form an important subring of K.

Definition 9.2.13 (Ring of Integers). The *ring of integers* in K is

$$\mathcal{O}_K = \{x \in K \mid x \text{ is integral over } \mathbb{Z}\}.$$

It's not at all clear from the definition just what $\mathcal{O}_K$ is, or even that it's a ring. Proposition 9.2.16 below will give a more explicit description of $\mathcal{O}_K$.

Lemma 9.2.14. *The map $K \to K$ given by*

$$a + b\sqrt{D} \mapsto \overline{a + b\sqrt{D}} = a - b\sqrt{D}$$

is an isomorphism of fields.

Proof. We have

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

□

Lemma 9.2.15. *Let $\alpha \in K$. The determinant of left multiplication by α on the 2-dimensional $\mathbb{Q}$-vector space K is $\mathrm{N}(\alpha) = \alpha\overline{\alpha}$. The trace of left multiplication is $\mathrm{Tr}(\alpha) = \alpha + \overline{\alpha}$. The characteristic polynomial of left multiplication by α is $x^2 - \mathrm{Tr}(\alpha)x + \mathrm{N}(\alpha)$.*

Proof. A basis for K as a $\mathbb{Q}$-vector space is $1, \sqrt{D}$. The matrix of left multiplication by $\alpha = a + b\sqrt{D}$ on this basis is $\begin{pmatrix} a & Db \\ b & a \end{pmatrix}$. Since $T(\alpha) = 2a$ and $\mathrm{N}(\alpha) = a^2 - Db^2$, the lemma follows. □

Proposition 9.2.16. *If $D \equiv 1 \pmod 4$ let $\alpha = (1+\sqrt{D})/2$, and otherwise let $\alpha = \sqrt{D}$. Then*

$$\mathcal{O}_K = \mathbb{Z}[\alpha] = \{a + b\alpha : a, b \in \mathbb{Z}\}.$$

Proof. First we prove that if $x = a + b\alpha \in \mathbb{Z}[\alpha]$, then $x \in \mathcal{O}_K$. By Lemma 9.2.15 it suffices to show that $\mathrm{Tr}(x)$ and $\mathrm{N}(x)$ lie in $\mathbb{Z}$. First we verify this for $x = \alpha$ by noting that $\mathrm{Tr}(\alpha) = 1$ and

$$\mathrm{N}(\alpha) = \begin{cases} (1 - D)/4 & \text{if } D \equiv 1 \pmod 4 \\ D & \text{otherwise.} \end{cases}$$

More generally, if $x = a + b\alpha$ with $a, b \in \mathbb{Z}$, then

$$\operatorname{Tr}(x) = \operatorname{Tr}(a + b\alpha) = 2a + b\operatorname{Tr}(\alpha) = 2a + b \in \mathbb{Z},$$

and

$$\begin{aligned} \mathrm{N}(x) &= (a + b\alpha)(a + b\overline{\alpha}) \\ &= a^2 + ab(\alpha + \overline{\alpha}) + b^2\alpha\overline{\alpha} \\ &= a^2 + ab\operatorname{Tr}(\alpha) + b^2\,\mathrm{N}(\alpha) \in \mathbb{Z}. \end{aligned}$$

For the other inclusion, suppose $x = a+b\sqrt{D} \in \mathbb{Q}(\sqrt{D})$ is integral over $\mathbb{Z}$. Then $\operatorname{Tr}(x) = 2a \in \mathbb{Z}$ and $\mathrm{N}(x) = a^2 - b^2D \in \mathbb{Z}$. Thus $a = a'/2$ for some $a' \in \mathbb{Z}$ and $(a')^2/4 - b^2D \in \mathbb{Z}$. Thus $(2b)^2D \in \mathbb{Z}$, so since D is square free the denominator of b is either 1 or 2. The denominator of b is 2 if and only if the denominator of a is 2 since $a^2 - b^2D \in \mathbb{Z}$. If the denominator of b is 1, then $a, b \in \mathbb{Z}$ and we are done, and if the denominator of b is 2, then $2b \in \mathbb{Z}$ and $(a')^2 \equiv (2b)^2D \pmod 4$, so D is a perfect square modulo 4 and hence $D \equiv 1 \pmod 4$ (since D is square free) and $x \in \mathbb{Z}[\alpha]$. □

Definition 9.2.17 (Field Discriminant). Let γ_1, γ_2 be any basis for $\mathcal{O}_K$, e.g., $\gamma_1 = 1$, $\gamma_2 = \alpha$ where α is as in Proposition 9.2.16. The *discriminant* of $\mathcal{O}_K$ is

$$d = \operatorname{disc}(\mathcal{O}_K) = \det\left(\begin{pmatrix} \gamma_1 & \gamma_1' \\ \gamma_2 & \gamma_2' \end{pmatrix}\right)^2.$$

Making a different choice of basis γ_1, γ_2 amounts to changing the determinant in the definition by ± 1, so the discriminant is well defined.

Proposition 9.2.18. *Let $K = \mathbb{Q}[\sqrt{D}]$ with D squarefree. Then the discriminant of K is D if $D \equiv 1 \pmod 4$ and $4D$ otherwise.*

Proof. First suppose $D \equiv 1 \pmod 4$. Then 1 and $\alpha = (1 + \sqrt{D})/2$ are a basis for $\mathcal{O}_K$, so

$$\begin{aligned} d &= \det\left(\begin{pmatrix} 1 & \alpha \\ 1 & \alpha' \end{pmatrix}\right)^2 \\ &= (-\alpha' - \alpha)^2 \\ &= (-\sqrt{D})^2 = D. \end{aligned}$$

On the other hand, if $D \not\equiv 1 \pmod 4$, then $\alpha = \sqrt{D}$ and $d = (-\sqrt{D} - \sqrt{D})^2 = 4D$. □

Let d be the discriminant of $\mathbb{Q}[\sqrt{D}]$. In Section 9.5 we will see that there is an bijection between certain equivalence classes of ideals in $\mathcal{O}_K$ and equivalence classes of binary quadratic forms of discriminant d. The set of equivalence classes of ideals in $\mathcal{O}_K$ will have the structure of finite abelian group induced by multiplication of ideals. Understanding whether or not numbers are represented by certain quadratic forms, is related to deciding whether or not certain ideals are principal in $\mathcal{O}_K$; this leads to class field theory, one of the major accomplishments of 20th century number theory.

9.3 Reduction Theory

Recall that a binary quadratic form is a polynomial of the form $f(x, y) = ax^2+bxy+cy^2$. Our motivating problem is to decide which numbers are represented by f; i.e., for which integers n do there exist integers x, y such that $ax^2+bxy+cy^2 = n$? If $g \in \mathrm{SL}_2(\mathbb{Z})$ then $f(x, y)$ and $f|_g(x, y) = f\left(g\begin{pmatrix}x\\y\end{pmatrix}\right)$ represent exactly the same set of integers. Also, $\mathrm{disc}(f) = \mathrm{disc}(f|_g)$, where $\mathrm{disc}(f) = b^2-4ac$, and f is called *positive definite* if $\mathrm{disc}(f) < 0$ and $a > 0$.

This section is about reduction theory, which allows us to decide whether or not two positive definite binary quadratic forms are equivalent under the action of $\mathrm{SL}_2(\mathbb{Z})$.

9.3.1 Reduced Forms

Definition 9.3.1 (Reduced). A positive definite quadratic form (a, b, c) is *reduced* if $|b| \leq a \leq c$ and if, in addition, when one of the two inequalities is an equality (i.e., either $|b| = a$ or $a = c$), then $b \geq 0$.

There is a geometric interpretation of the notion of being reduced. Let $D = \mathrm{disc}(a, b, c) = b^2 - 4ac$ and set $\tau = \frac{-b+\sqrt{D}}{2a}$, so τ is the root of $ax^2 + bx + c$ with positive imaginary part. The right action of $\mathrm{SL}_2(\mathbb{Z})$ on positive definite binary quadratic forms corresponds to the left action of $\mathrm{SL}_2(\mathbb{Z})$ by linear fractional transformations on the complex upper half plane $\mathfrak{h} = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$. The standard "fundamental domain" for the action of $\mathrm{SL}_2(\mathbb{Z})$ on $\mathfrak{h}$ is

$$\mathcal{F} = \left\{\tau \in \mathfrak{h} : \mathrm{Re}(\tau) \in \left[-\frac{1}{2}, \frac{1}{2}\right), |\tau| > 1 \text{ or } |\tau| = 1 \text{ and } \mathrm{Re}(\tau) \leq 0\right\}.$$

Then (a, b, c) is reduced if and only if the corresponding complex number τ lies in $\mathcal{F}$. For example, if (a, b, c) is reduced then $\mathrm{Re}(\tau) = -b/2a \in [-1/2, 1/2)$ since $|b| \leq a$ and if $|b| = a$ then $b \geq 0$. Also

$$|\tau| = \sqrt{\frac{b^2 + 4ac - b^2}{4a^2}} = \sqrt{\frac{c}{a}} \geq 1$$

and if $|\tau| = 1$ then $b \geq 0$ so $\mathrm{Re}(\tau) \leq 0$.

The following theorem highlights the importance of reduced forms.

Theorem 9.3.2. *There is exactly one reduced form in each equivalence class of positive definite binary quadratic forms.*

Proof. We have to prove two things. First, that every class contains at least one reduced form, and second that this reduced form is the only one in the class.

We first prove that there is a reduced form in every class. Let $\mathcal{C}$ be an equivalence class of positive definite quadratic forms of discriminant D. Let (a, b, c) be an element of $\mathcal{C}$ such that a is minimal (amongst elements of $\mathcal{C}$). Note that for any such form we have $c \geq a$, since (a, b, c) is equivalent to $(c, -b, a)$ (use the matrix $\left(\begin{smallmatrix}0 & -1\\ 1 & 0\end{smallmatrix}\right)$). Applying the element $\left(\begin{smallmatrix}1 & k\\ 0 & 1\end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$ to

(a, b, c) for a suitably chosen integer k (precisely, $k = \lfloor (a-b)/2a \rfloor$) results in a form (a', b', c') with $a' = a$ and $b' \in (-a', a']$. Since $a' = a$ is minimal, we have just as above that $a' \leq c'$, hence (a', b', c') is "just about" reduced. The only possible remaining problem would occur if $a' = c'$ and $b' < 0$. In that case, changing (a', b', c') to $(c'', b'', a'') = (c', -b', a')$ results in an equivalent form with $b'' > 0$, so that (c'', b'', a'') is reduced.

Next suppose (a, b, c) is a reduced form. We will now establish that (a, b, c) is the only reduced form in its equivalence class. First, we check that a is minimal amongst all forms equivalent to (a, b, c). Indeed, every other a' has the form $a' = ap^2 + bpr + cr^2$ with p, r coprime integers (see this by hitting (a, b, c) by $\left(\begin{smallmatrix} p & q \\ r & s \end{smallmatrix}\right)$). The identities

$$ap^2 + bpr + cr^2 = ap^2\left(1 + \frac{b}{a}\frac{r}{p}\right) + cr^2 = ap^2 + cr^2\left(1 + \frac{b}{c}\frac{p}{r}\right)$$

then imply our claim since $|b| \leq a \leq c$ (use the first identity if $r/p < 1$ and the second otherwise). Thus any other reduced form (a', b', c') equivalent to (a, b, c) has $a' = a$. But the same identity implies that the only forms equivalent to (a, b, c) with $a' = a$ are obtained by applying a transformation of the form $\begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}$ (corresponding to $p = 1$, $r = 0$). Thus $b' = b + 2ak$ for some k. Since $a = a'$ we have $b, b' \in (-a, a]$, so $k = 0$. Finally

$$c' = \frac{(b')^2 - D}{4a'} = \frac{b^2 - D}{4a} = c,$$

so $(a', b', c') = (a, b, c)$. □

9.3.2 *Finding an Equivalent Reduced Form*

Here is how to find the reduced form equivalent to a given positive definite form (a, b, c). This algorithm is useful for solving problems 8 and 9 on the homework assignment. Consider the following two operations, which can be used to diminish one of a and $|b|$, without altering the other:

1. If $c < a$, replace (a, b, c) by the equivalent form $(c, -b, a)$.

2. If $|b| > a$, replace (a, b, c) by the equivalent form (a, b', c') where $b' = b + 2ka$ and k is chosen so that $b' \in (-a, a]$ (more precisely, $k = \lfloor \frac{a-b}{2a} \rfloor$), and c' is found from the fact that $(b')^2 - 4ac' = D = \mathrm{disc}(a, b, c)$, so $c' = \frac{(b')^2 - D}{4a}$.

Starting with (a, b, c), if you iterate the appropriate operation, eventually you will find the reduced form that is equivalent to (a, b, c).

Example 9.3.3. Let $f = 458x^2 + 214xy + 25y^2$.

Equivalent form	What I did	Matrix
$(458, 214, 25)$		
$(25, -214, 458)$	(1)	$\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$
$(25, -14, 2)$	(2) with $k = 4$	$\left(\begin{smallmatrix} 1 & 4 \\ 0 & 1 \end{smallmatrix}\right)$
$(2, 14, 25)$	(1)	$\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$
$(2, 2, 1)$	(2) with $k = -3$	$\left(\begin{smallmatrix} 1 & -3 \\ 0 & 1 \end{smallmatrix}\right)$
$(1, -2, 2)$	(1)	$\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$
$(1, 0, 1)$	(2) with $k = 1$	$\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$

Let

$$\begin{aligned} g &= \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 4 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & -3 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 3 & 4 \\ -13 & -17 \end{pmatrix}. \end{aligned}$$

Then

$$f|_g = x^2 + y^2!$$

Example 9.3.4. If f_1 and f_2 are binary quadratic forms that represent exactly the same integers, is $f_1 \sim f_2$? The answer is no. For example, $f_1 = (2, 1, 3) = 2x^2 + xy + 3y^2$ and $f_2 = (2, -1, 3) = 2x^2 - xy + 3y^2$ are inequivalent reduced positive definite binary quadratic forms that represent exactly the same integers. Note that $\mathrm{disc}(f_1) = \mathrm{disc}(f_2) = -23$.

9.4 Class Numbers

Proposition 9.4.1. *Let $D < 0$ be a discriminant. There are only finitely many equivalence classes of positive definite binary quadratic forms of discriminant D.*

Proof. Since there is exactly one reduced binary quadratic form in each equivalence class, it suffices to show that there are only finitely many reduced forms of discriminant D. Recall that if a form (a, b, c) is reduced, then $|b| \leq a \leq c$. If (a, b, c) has discriminant D then $b^2 - 4ac = D$. Since $b^2 \leq a^2 \leq ac$, we have $D = b^2 - 4ac \leq -3ac$, so

$$3ac \leq -D.$$

There are only finitely many positive integers a, c that satisfy this inequality. □

Definition 9.4.2. A binary quadratic form (a, b, c) is called *primitive* if $\gcd(a, b, c) = 1$.

Definition 9.4.3. The *class number* h_D of discriminant $D < 0$ is the number of equivalence classes of primitive positive definite binary quadratic forms of discriminant D.

Table 9.1 lists the class numbers h_D for $-D \leq 599$ with D odd. Notice that there are just a few 1s at the beginning and then no more.

TABLE 9.1. Class Numbers For D Odd

$-D$	h_D	$-D$	h_D	$-D$	h_D	$-D$	h_D	$-D$	h_D
3	**1**	123	2	243	3	363	4	483	4
7	**1**	127	5	247	6	367	9	487	7
11	**1**	131	5	251	7	371	8	491	9
15	2	135	6	255	12	375	10	495	16
19	**1**	139	3	259	4	379	3	499	3
23	3	143	10	263	13	383	17	503	21
27	**1**	147	2	267	2	387	4	507	4
31	3	151	7	271	11	391	14	511	14
35	2	155	4	275	4	395	8	515	6
39	4	159	10	279	12	399	16	519	18
43	**1**	163	**1**	283	3	403	2	523	5
47	5	167	11	287	14	407	16	527	18
51	2	171	4	291	4	411	6	531	6
55	4	175	6	295	8	415	10	535	14
59	3	179	5	299	8	419	9	539	8
63	4	183	8	303	10	423	10	543	12
67	**1**	187	2	307	3	427	2	547	3
71	7	191	13	311	19	431	21	551	26
75	2	195	4	315	4	435	4	555	4
79	5	199	9	319	10	439	15	559	16
83	3	203	4	323	4	443	5	563	9
87	6	207	6	327	12	447	14	567	12
91	2	211	3	331	3	451	6	571	5
95	8	215	14	335	18	455	20	575	18
99	2	219	4	339	6	459	6	579	8
103	5	223	7	343	7	463	7	583	8
107	3	227	5	347	5	467	7	587	7
111	8	231	12	351	12	471	16	591	22
115	2	235	2	355	4	475	4	595	4
119	10	239	15	359	19	479	25	599	25

Theorem 9.4.4 (Heegner, Stark-Baker, Goldfeld-Gross-Zagier). *Suppose D is a negative fundamental discriminant. Then*

- $h_D = 1$ *only for* $D = -3, -4, -7, -8, -11, -19, -43, -67, -163$.
- $h_D = 2$ *only for* $D = -15, -20, -24, -35, -40, -51, -52, -88, -91, -115, -123, -148, -187, -232, -235, -267, -403, -427$.
- $h_D = 3$ *only for* $D = -23, -31, -59, -83, -107, -139, -211, -283, -307, -331, -379, -499, -547, -643, -883, -907$.
- $h_D = 4$ *only for* $D = -39, -55, -56, -68, \ldots, -1555$.

To quote Henri Cohen: "The first two statements concerning class numbers 1 and 2 are very difficult theorems proved in 1952 by Heegner and in 1968–1970 by Stark and Baker. The general problem of determing all imaginary quadratic fields with a given class number has been solved in principle by Goldfeld-Gross-Zagier, but to my knowledge the explicit computations have been carried to the end only for class numbers 3 and 4 (in addition to the already known class numbers 1 and 2).

This is somewhat out of date[4] from Mathworld... For what it's worth: 4

```
http://mathworld.wolfram.com/GausssClassNumberProblem.html

``Oesterl\'e (1985) solved the case h=3,
and Arno (1992) solved the case h=4.  Wagner (1996) solve[d]
the cases n=5, 6, 7.  Arno et al. (1993) solved the problem
for odd h satisfying 5 \leq h \leq 23.  In his thesis,
M. Watkins has solved the problem for all h \leq 16."
```

[4] Finish later.

9.5 Correspondence Between Binary Quadratic Forms and Ideals

In this section we describe a bijection between certain equivalence classes of ideals and certain equivalence classes of binary quadratic forms. Since equivalence classes of ideals have a natural group structure, this bijection induces a group structure on equivalence classes of binary quadratic forms.

For the rest of this section, $K = \mathbb{Q}(\sqrt{d})$ is a quadratic field with discriminant d (see Definition 9.2.17). Thus $d \equiv 1 \pmod 4$ and d is square free, or $d \equiv 0 \pmod 4$ and $d/4$ is square free and $d/4 \not\equiv 1 \pmod 4$. Let $\mathcal{O}_K$ denote the ring of all algebraic integers in K, as in Section 9.2.5.

9.5.1 Correctly Ordered Basis For Ideals

Proposition 9.5.1. *Suppose $I \subset \mathcal{O}_K$ is an ideal. Then there exists $\alpha, \beta \in \mathcal{O}_K$ such that*

$$I = \mathbb{Z}\alpha + \mathbb{Z}\beta = \{x\alpha + y\beta : x, y \in \mathbb{Z}\}.$$

Proof. As an abelian group, $\mathcal{O}_K$ is isomorphic to $\mathbb{Z}^2$. By the structure theorem for finitely generated abelian groups and the fact that subgroups of finitely generated abelian groups are finitely generated, I is isomorphic to $\mathbb{Z}^r$ for some positive integer r. Thus there is an inclusion $\mathbb{Z}^r \to \mathbb{Z}^2$. This extends to an injective vector space homomorphism $\mathbb{Q}^r \to \mathbb{Q}^2$, so $r \leq 2$. Since I is an ideal, I has finite index in $\mathcal{O}_K$ (see Exercise 6), so $r \geq 2$. Thus I is generated as a $\mathbb{Z}$-module by two elements, α and β. $\square$

We view $[\alpha, \beta]$ as remembering our choice of ordered basis α, β. When used as an ideal, interpret $[\alpha, \beta]$ to mean $\mathbb{Z}\alpha + \mathbb{Z}\beta$. Thus Proposition 9.5.1 asserts that for every ideal I is of the form $[\alpha, \beta]$ for some $\alpha, \beta \in \mathcal{O}_K$. Note, however, that there are many choices of α, β so that $[\alpha, \beta]$ is not an ideal. For example, $[2, i]$ in $\mathbb{Z}[i]$ is not equal to $\mathbb{Z}[i]$, but it contains the unit i, so if it were an ideal then it would have to equal $\mathbb{Z}[i]$.

It is natural to define a binary quadratic form associated to $I = [\alpha, \beta]$ as follows:

$$\begin{aligned} Q &= \mathrm{N}(\alpha x + \beta y) \\ &= (\alpha x + \beta y)(\alpha' x + \beta' y) \\ &= (\alpha\alpha')x^2 + (\alpha\beta' + \beta\alpha')xy + \beta\beta' y^2. \end{aligned}$$

Surprisingly, this definition would lead to a disasterous breakdown of the theory! The quadratic form associated to I and the conjugate $I' = [\alpha', \beta']$ would be the same. In Section 9.5.3 we will define a group structure on certain equivalence classes of ideals, and in this group the equivalence classes $[I]$ and $[I']$ are inverses. Since there should be a bijection between equivalence classes of binary quadratic forms and ideal classes, we would have to have that $[I][I] = [I][I'] = 1$, so the group of ideals would be a finite 2-torsion group, hence have order a power of 2, which is generally not the case. We must be much more careful in how we associate a binary quadratic form to an ideal.

Definition 9.5.2 (Correctly Ordered). A basis $[\alpha, \beta]$ for an ideal I is *correctly ordered* if

$$\frac{\alpha\beta' - \beta\alpha'}{\sqrt{d}} > 0.$$

Example 9.5.3. Let $d = -4$ and let I be the ideal generated by $(5, i-2)$. Note that $\mathcal{O}_K = \mathbb{Z}[i]$, so $\mathcal{O}_K/I \cong \mathbb{Z}/5$. We have $I = [5, i-2]$, since $[5, i-2] \subset I$ and $\det\left(\begin{smallmatrix} 5 & -2 \\ 0 & 1 \end{smallmatrix}\right) = 5$ (so that $\#(\mathcal{O}_K/[5, i-2]) = 5$). Notice that $[5, i-2]$ is not correctly ordered because

$$\frac{5(-i-2) - (i-2)5}{\sqrt{-4}} = -5 < 0.$$

The basis $[i-2, 5]$ is correctly ordered.

Proposition 9.5.4. *Any two correctly ordered bases of an ideal I are equivalent by an element in* $\mathrm{SL}_2(\mathbb{Z})$*, and conversely.*

Proof. Suppose $[\alpha, \beta] = [\gamma, \delta]$ are two correctly ordered basis for an ideal I. Because these are two different basis for the same free $\mathbb{Z}$-module, there are $a, b, c, d \in \mathbb{Z}$ such that

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \gamma \\ \delta \end{pmatrix} = A \begin{pmatrix} \gamma \\ \delta \end{pmatrix},$$

and $\det(A) = \pm 1$ (this is just like a change of basis matrix in linear algebra; its determinant is a unit in the base ring). Since $a, b, c, d \in \mathbb{Z}$ and the conjugation automorphism fixes $\mathbb{Z}$, we have

$$\begin{pmatrix} \alpha & \alpha' \\ \beta & \beta' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \gamma & \gamma' \\ \delta & \delta' \end{pmatrix}.$$

Taking determinants, we have

$$\alpha\beta' - \beta\alpha' = \det(A)(\gamma\delta' - \delta\gamma'). \tag{9.1}$$

Since $[\alpha, \beta]$ and $[\gamma, \delta]$ are correctly oriented, we must have $\det(A) = +1$, so $A \in \mathrm{SL}_2(\mathbb{Z})$.

Conversely, if $A \in \mathrm{SL}_2(\mathbb{Z})$ and $[\gamma, \delta]$ is a correctly oriented basis then,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \gamma & \gamma' \\ \delta & \delta' \end{pmatrix} = \begin{pmatrix} \alpha & \alpha' \\ \beta & \beta' \end{pmatrix}$$

and by (9.1) $[\alpha, \beta]$ is also correctly oriented. □

9.5.2 Norms of Ideals

Definition 9.5.5 (Norm). The *norm* of a nonzero ideal I of $\mathcal{O}_K$ is the positive integer

$$\mathrm{N}(I) = \#(\mathcal{O}_K/I).$$

Proposition 9.5.6. $II' = (N(I))$

Proof. A complete proof is given on page 128-129 of Cohn "Advanced Number Theory". This fact follows from "Hurwitz's Lemma", i.e., it is nontrivial. [5] $\square$ 5

Lemma 9.5.7. *Let (a, b) and (c, d) be elements of $\mathbb{Z} \oplus \mathbb{Z}$ such that $D = |\det\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)| \neq 0$. Then the quotient abelian group*

$$M = (\mathbb{Z} \oplus \mathbb{Z})/(\mathbb{Z}(a, b) + \mathbb{Z}(c, d))$$

is finite of order D.

Proof. Let $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. By repeatedly swapping rows, swapping columns, adding a multiple of one row to another row, or adding a multiple of one column to another column, we can transform A into a diagonal matrix $\left(\begin{smallmatrix} e & 0 \\ 0 & f \end{smallmatrix}\right)$. Each swapping and adding operations changes at most change the sign of the determinant, so $|\det(A)| = |ef|$. We may thus assume that A is diagonal, in which case the lemma is clear. $\square$

Lemma 9.5.8. *Suppose I is an ideal of $\mathcal{O}_K$ with basis $[\alpha, \beta]$, and let d be the discriminant of K. Then*

$$\det\begin{pmatrix} \alpha & \alpha' \\ \beta & \beta' \end{pmatrix}^2 = d \cdot \mathrm{N}(I)^2.$$

Proof. Let γ_1, γ_2 be a basis for $\mathcal{O}_K$. Since α and β can be written as a $\mathbb{Z}$-linear combination of γ_1 and γ_2 there is a 2×2 integer matrix A such that

$$A\begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

We have

$$\begin{aligned} \det\begin{pmatrix} \alpha & \alpha' \\ \beta & \beta' \end{pmatrix}^2 &= \det\left(A \cdot \begin{pmatrix} \gamma_1 & \gamma_1' \\ \gamma_2 & \gamma_2' \end{pmatrix}\right)^2 \\ &= \det(A)^2 d \\ &= \mathrm{N}(I)^2 d, \end{aligned}$$

where we use Lemma 9.5.7 to see that $\det(A) = \mathrm{N}(I)$. $\square$

9.5.3 The Ideal Class Group

For the rest of this section, let $K = \mathbb{Q}(\sqrt{d})$ be a quadratic field with discriminant d.

Definition 9.5.9. Two ideals $I, J \subset \mathcal{O}_K$ are *equivalent*, written $I \sim J$, if there are $\alpha, \beta \in \mathcal{O}_K$ such that

$$\alpha I = \beta J \qquad \text{and} \qquad \mathrm{N}(\alpha\beta) > 0.$$

We will denote the set of equivalence classes of **nonzero** ideals in $\mathcal{O}_K$ by $\mathrm{Cl}^+(\mathcal{O}_K)$.

---

[5] Write a proof for the notes based on Cohn's proof.

Proposition 9.5.10. *Multiplication of ideals gives* $\mathrm{Cl}^+(\mathcal{O}_K)$ *an abelian group structure in which the ideal class of* $\mathcal{O}_K = (1)$ *is the identity element.*

Proof. If I and J are ideals of $\mathcal{O}_K$ then their product $IJ = \{xy : x \in I, y \in J\}$ is again an ideal in $\mathcal{O}_K$. Multiplication of ideals is easily seen to be associative and commutative, since the usual multiplication of elements in $\mathcal{O}_K$ is associative and commutative. Next we check that multiplication of ideals induces a well-defined multiplication of ideal classes. If $I_0 \sim J_0$ and $I_1 \sim J_1$, then there exists $\alpha_0, \beta_0, \alpha_1, \beta_1 \in \mathcal{O}_K$ such that $\alpha_i I_i = \beta_i J_i$ for $i = 0, 1$. Multiplying these two equalities, we see that $\alpha_0 I_0 \alpha_1 I_1 = \beta_0 J_0 \beta_1 J_1$, so $\alpha_0\alpha_1 I_0 I_1 = \beta_0\beta_1 J_0 J_1$, hence $I_0 I_1 \sim J_0 J_1$. This proves that multiplication of ideals induces a well-defined associative commutative multiplication of ideal classes.

To finish the proof, we verify that every element of $\mathrm{Cl}^+(\mathcal{O}_K)$ has an inverse. Let $I = [\alpha, \beta]$ be an ideal in $\mathcal{O}_K$. Then the ideal I' generated by the conjugates of elements of I is $[\alpha', \beta']$. By Proposition 9.5.6, the product II' is the principal ideal generated by the positive integer $\#(\mathcal{O}_K/I)$. Thus $II' \sim (1)$, so I has an inverse. □

Example 9.5.11. Let $K = \mathbb{Q}[\sqrt{-20}]$. Then $\mathrm{Cl}^+(\mathcal{O}_K)$ is cyclic of order 2. A nonidentity element of $\mathrm{Cl}^+(\mathcal{O}_K)$ is $I = [1 + \sqrt{-5}, 2]$.

9.5.4 Correspondence Between Ideals and Forms

Recall that a binary quadratic form $Q = ax^2 + bxy + cy^2$ is primitive if $\gcd(a, b, c) = 1$ and has discriminant $b^2 - 4ac$. The following proposition associates a primitive binary quadratic form to an ideal of $\mathcal{O}_K$.

Proposition 9.5.12. *Let* I *be an ideal in* $\mathcal{O}_K$ *and let* $[\alpha, \beta]$ *be a correctly ordered basis for* I*. Then the quadratic form*

$$Q = \frac{\mathrm{N}(\alpha x + \beta y)}{\mathrm{N}(I)} = ax^2 + bxy + cy^2$$

has integral coefficients and is a primitive form of discriminant d*.*

Note that the numerator $\mathrm{N}(\alpha x + \beta y)$ in the definition of Q depends on the order of α and β.

Proof. We have

$$\begin{aligned} \mathrm{N}(\alpha x + \beta y) &= (\alpha x + \beta y)(\alpha' x + \beta' y) \\ &= \alpha\alpha' x^2 + (\alpha\beta' + \alpha'\beta)xy + \beta\beta' y^2 \\ &= Ax^2 + Bxy + Cy^2. \end{aligned}$$

The coefficients A, B, and C are elements of $\mathbb{Z}$ because they are norms and traces. They are also elements of $(\mathrm{N}(I))$, since they are visibly elements of II' and by Proposition 9.5.6, $II' = (\mathrm{N}(I))$. Thus there exists $a, b, c \in \mathcal{O}_K$ such that

$$\begin{array}{rclcl} A &=& \alpha\alpha' &=& a\,\mathrm{N}(I), \\ B &=& \alpha\beta' + \alpha'\beta &=& b\,\mathrm{N}(I), \\ C &=& \beta\beta' &=& c\,\mathrm{N}(I). \end{array}$$

Since A and $\mathrm{N}(I)$ are both in $\mathbb{Z}$ and $a \in \mathcal{O}_K$, we see that $a \in \mathbb{Z}$; likewise, $b, c \in \mathbb{Z}$. Thus $Q = ax^2 + bxy + cy^2$ has coefficients in $\mathbb{Z}$.

By Lemma 9.5.8,

$$\begin{aligned} b^2 - 4ac &= (B^2 - 4AC)/\,\mathrm{N}(I)^2 \\ &= (\alpha\beta' - \beta\alpha')^2/\,\mathrm{N}(I)^2 = d, \end{aligned}$$

where d is the discriminant of K.

All that remains is to show that $\gcd(a, b, c) = 1$. If f is a positive divisor of $\gcd(a, b, c)$, then $f^2 \mid b^2 - 4ac = d$. If $d \equiv 1 \pmod 4$ then d is square free so $f = 1$. If $d \equiv 0 \pmod 4$ then $d' = d/4$ is square free and $d' \not\equiv 1 \pmod 4$, so $f = 1$ or $f = 2$. If $f = 2$ write $a = 2a'$, $b = 2b'$, and $c = 2c'$ for integers a', b', c' with b' odd. Then

$$b^2 - 4ac = 4b'^2 - 16a'c' = 4d'.$$

Reducing this equation modulo 16 implies that $4b'^2 \equiv 4d' \pmod{16}$. Dividing this congruence through by 4 implies that $b'^2 \equiv d' \pmod 4$. Since b' is odd, $b'^2 \equiv 1 \pmod 4$, which contradicts the fact that $d' \not\equiv 1 \pmod 4$. Thus $f = 1$ in all cases, so $ax^2 + bxy + cy^2$ is primitive. □

Example 9.5.13. Let $K = \mathbb{Q}[\sqrt{-20}]$ and $I = [1 + \sqrt{-5}, 2]$, as in Example 9.5.11. Then

$$\begin{aligned} \mathrm{N}(\alpha x + \beta y) &= ((1 + \sqrt{-5})x + 2y)((1 - \sqrt{-5})x + 2y) \\ &= 6x^2 + 4xy + 4y^2. \end{aligned}$$

The norm of I is $|\det(\left(\begin{smallmatrix} 1 & 2 \\ 1 & 0 \end{smallmatrix}\right))| = 2$. Thus $Q = 3x^2 + 2xy + 2y^2$. Notice, as a check, that $\mathrm{disc}(Q) = b^2 - 4ac = 2^2 - 4 \cdot 3 \cdot 2 = -20$.

Example 9.5.14. Let $K = \mathbb{Q}[\sqrt{23}]$, which has discriminant $d = 92$. The ideal $I = (\sqrt{23})$ is principal, but it is not equivalent to $(1) \in \mathrm{Cl}^+(\mathcal{O}_K)$. The quadratic form associated to $I = [\sqrt{23}, 23]$ is

$$Q = \frac{-23x^2 + 23^2y^2}{23} = -x^2 + 23y^2.$$

The quadratic form associated to $(1) = [\sqrt{23}, 1]$ is $R = -23x^2 + y^2$. These two forms can not be equivalent since Q represents -1 but R doesn't (since modulo 4 we have $R \equiv x^2 + y^2$, which never takes on the value $3 \mod 4$).

MAGMA **incorrectly asserts that Q and R are equivalent. I sent in a bug report this morning.**

Proposition 9.5.15. *Let $Q = ax^2 + bxy + cy^2$ be a primitive binary quadratic form of discriminant d (if $d < 0$ assume that $a > 0$). Then*

$$I = [\alpha, \beta] = \begin{cases} [a, \frac{b - \sqrt{d}}{2}] & \text{if } a > 0, \\ [a\sqrt{d}, (\frac{b - \sqrt{d}}{2})\sqrt{d}] & \text{if } a < 0. \end{cases}$$

is an ideal of $\mathcal{O}_K$ and $[\alpha, \beta]$ is a correctly ordered basis for I.

Example 9.5.16. Let $d = -20$ and $Q = 3x^2 + 2xy + 2y^2$. Then $I = [3, 1 - \sqrt{-5}]$. Notice that $I \neq [1 + \sqrt{-5}, 2]$, so the operations of the two propositions are not inverses before passing to equivalence classes.

Proof. All we have to do is check that $\gamma = (b - \sqrt{d})/2$ is in $\mathcal{O}_K$ and that I is correctly ordered. If $d \equiv 1 \pmod 4$ then b is odd, so $\gamma \in \mathcal{O}_K = \mathbb{Z}[(1+\sqrt{d})/2]$, and if $d \equiv 0 \pmod 4$ then b is even, so

$$\gamma = \frac{2b' - 2\sqrt{d'/4}}{2} \in \mathcal{O}_K = \mathbb{Z}[\sqrt{d'}].$$

It is straightforward but tedious to check that the given basis is ordered.
[6] □ 6

Theorem 9.5.17. *Let K be a quadratic field with discriminant d. Let $\mathcal{Q}(d)$ be the set of equivalence classes of primitive binary quadratic forms of discriminant d (if $d < 0$ include only positive definite forms in $\mathcal{Q}$). Then Propositions 9.5.12 and 9.5.15 induce a bijection between $\mathcal{Q}(d)$ and $\mathrm{Cl}^+(\mathcal{O}_K)$. In particular, $\mathcal{Q}(d)$ has the structure of finite abelian group.*

The proof is on pages 204–206 of Cohn's book. It doesn't look difficult, but it is long and tedious, so we will omit it. [7] 7

[6] Add details.

[7] But maybe add it to the book.

Exercises

9.1 Which of the following numbers is a sum of two squares? Express those that are as a sum of two squares.

$$-389,\ 12345,\ 91210,\ 729,\ 1729,\ 68252$$

9.2 (a) Write a simple computer program that takes a positive integer n as input and outputs a sequence `[x,y,z,w]` of four integers such that $x^2 + y^2 + z^2 + w^2 = n$.

(b) Write 2001 as a sum of three squares.

9.3 Find a positive integer that has at least three different representations as the sum of two squares, disregarding signs and the order of the summands.

9.4 Show that a natural number n is the sum of two integer squares if and only if it is the sum of two rational squares.

9.5 Mimic the proof of the main theorem of Lecture 21 to show that an odd prime p is of the form $8m + 1$ or $8m + 3$ if and only if it can be written as $p = x^2 + 2y^2$ for some choice of integers x and y.

9.6 Let K be a quadratic field and let I be a nonzero ideal in $\mathcal{O}_K$. Use Lemma 9.5.7 to prove the $\mathcal{O}_K/I$ is finite.

9.7 A *triangular number* is a number that is the sum of the first m integers for some positive integer m. If n is a triangular number, show that all three of the integers $8n^2$, $8n^2 + 1$, and $8n^2 + 2$ can be written as a sum of two squares.

9.8 Prove that of any four consecutive integers, at least one is not representable as a sum of two squares.

9.9 Show directly that $13x^2 + 36xy + 25y^2$ and $58x^2 + 82xy + 29y^2$ are each equivalent to the form $x^2 + y^2$, then find integers x and y such that $13x^2 + 36xy + 25y^2 = 389$.

9.10 What are the discriminants of the forms $199x^2 - 162xy + 33y^2$ and $35x^2 - 96xy + 66y^2$? Are these forms equivalent?

9.11 For any negative discriminant D, let C_D denote the finite abelian group of equivalence classes of primitive positive definite quadratic forms of discriminant D. Use a computer to compute representatives for C_D and determine the structure of C_D as a produce of cyclic groups for each of the following five values of D:

$$D = -155, -231, -660, -12104, -10015.$$

Part II

Elliptic Curves

10

Elliptic Curves and Their Groups

An elliptic curve is a nonsingular curve which can be defined by an equation the form $y^2 = x^3 + ax + b$ with $4a^3 + 27b^2 \neq 0$. Elliptic curves are one of the most exciting and central objects in modern number theory. Several very deep results were proved about elliptic curves in the last two decades, as we will see in Chapters 12, **??**, and 13.[1] 1

This chapter establishes the basic foundations on which subsequent chapters will build. We begin with the definition of an elliptic curve over the complex numbers in Section 10.1. Section **??** is about how they have a natural structure of abelian group. In Section **??** we return to number theory, and consider the subgroup of points with coordinates in a fixed field, such as $\mathbb{Q}$.

Though we will use the word "natural" many times in this chapter, the reader should not try to assign too precise a meaning to this word. Any nonempty set can be endowed with an abelian group structure (see Exercise 10), so it is no accomplish to prove that a set has a group structure, unless that group structure is in some way natural.

10.1 The Definition of Elliptic Curves over $\mathbb{C}$

This section is about how to define elliptic curves over the complex number $\mathbb{C}$.

Recall from the introduction to this chapter that an elliptic is a curve that can be "defined" by an equation of the form $y^2 = x^3 + ax + b$. This is technically correct, but sneaky, because the phrase "can be defined" is vague. The graph Y of $y^2 = x^3 + ax + b$ is the set of points in $\mathbb{C}^2 = \mathbb{C} \times \mathbb{C}$

[1] Expand and elaborate.

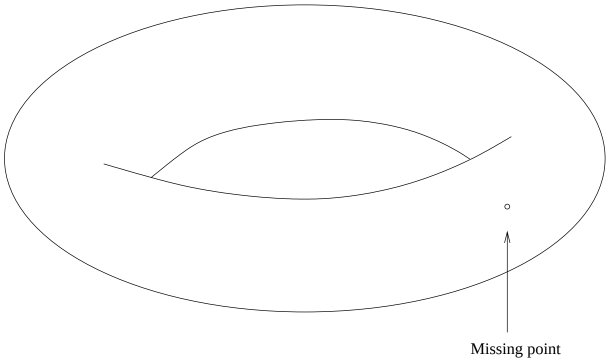


FIGURE 10.1. Y is Homeomorphic to a Torus With a Point Removed

that satisfy $y^2 = x^3 + ax + b$, so

$$Y = \{(z, w) \in \mathbb{C}^2 : w^2 = z^3 + az + b\} \subset \mathbb{C}^2.$$

Since for each value of z there are exactly two values of w that satisfy the equation (counting "multiplicity"), Y has dimension 2 when viewed as a real manifold (locally Y looks like a small open subset of $\mathbb{C}$). When viewed as a complex manifold, Y has dimension 1, which is one reason that we refer to Y as a "curve".

Using the Weierstrass $\wp$ function from complex analysis (which will appear again in Section 10.2.5) one can see that there is a homeomorphism between Y and a torus with one point removed (see Figure 10.1). Notice that Y is "incomplete" in the sense that it is visibly missing a point. The set Y is closed when viewed as a subset of $\mathbb{C}^2$, since it is the inverse image of 0 under the continuous map $\mathbb{C}^2 \to \mathbb{C}$ given by $(x, y) \mapsto y^2 - x^3 - ax - b$. Thus we won't find the missing point by taking the closure of Y in $\mathbb{C}^2$. We instead consider Y as a subset of the project plane, where the closure will include the one missing point.

10.1.1 Some Topology

Before defining the projective plane, we review some basic topology.

A *topological space* is a set X together with a collection of open subsets $U \subset X$ that are closed under arbitrary unions, finite intersections, and X and $\emptyset$ are open. A subset $A \subset X$ is *closed* if $X \setminus A$ is open.

A *continuous map* $f : X \to Y$ of topological spaces is a map such that whenever $U \subset Y$ is open, the inverse image $f^{-1}(U)$ is open in X. Because $f^{-1}(X \setminus A) = X \setminus f^{-1}(A)$, one sees at once that f is continuous if and only if the inverse image of every closed subset of Y is closed in X.

A subset $U \subset X$ is *clopen* if it is both open and closed. A topological space X is *connected* if it has no clopen subsets besides $\emptyset$ and X. The continuous image of a connected set is connected, because if $f : X \to Y$ is continuous and $C \subset Y$ is clopen, then $f^{-1}(C) \subset X$ is also clopen, and $f^{-1}(C) = X$ or $\emptyset$ if and only if $C = Y$ or $\emptyset$.

Suppose X is a topological space and $f : X \to Y$ is a map from X to a set Y. The *induced topology* on Y is the topology in which the open subsets of Y are the subsets $U \subset Y$ such that $f^{-1}(U)$ is an open subset of X. This is the "coarsest" topology on Y that makes the map f continuous.

We assume the reader is familiar with the standard topology on $\mathbb{R}$ and $\mathbb{C}$ from a course in Calculus or Analysis (open subsets of $\mathbb{R}$ are unions of open intervals (a, b), etc.).

10.1.2 The Projective Plane

Definition 10.1.1 (Projective Plane). The *projective plane* $\mathbb{P}^2$ is the set of triples $(a, b, c) \in \mathbb{C}^3$ with a, b, c not all 0, modulo the equivalence relation

$$(a, b, c) \sim (\lambda a, \lambda b, \lambda c)$$

for all nonzero $\lambda \in \mathbb{C}$. Denote by $(a : b : c)$ the equivalence class of (a, b, c). The topology on $\mathbb{P}^2$ is the one induced by viewing it as a quotient of $\mathbb{C}^3 \setminus \{0\}$.

The projective plane is a little bigger than the usual plane, in the following sense. There is a map $\mathbb{C}^2 \hookrightarrow \mathbb{P}^2$ that sends (a, b) to $(a : b : 1)$, and the complement of the image is a projective line:

$$\mathbb{P}^2 \setminus \mathbb{C}^2 \cong \{(a : 1 : 0) : a \in \mathbb{C}\} \cup \{(1 : 0 : 0)\}.$$

Thus $\mathbb{P}^2$ is set-theoretically the disjoint union

$$\mathbb{C}^2 \cup \mathbb{C} \cup \{\text{point}\}.$$

Since $\mathbb{P}^2$ is a continuous image of the connected set $\mathbb{C}^3 \setminus \{0\}$, we see that $\mathbb{P}^2$ is also connected, so we should not view $\mathbb{P}^2$ "topologically" as the above disjoint union. The inverse image of $\mathbb{C}^2$ in $\mathbb{C}^3 \setminus \{0\}$ is the complement of the (closed) plane $\{(a, b, 0) : a, b \in \mathbb{C}\}$, so $\mathbb{C}^2$ is an open subset of $\mathbb{P}^2$. Since $\mathbb{P}^2$ is connected, $\mathbb{C}^2$ is *not* a closed subset of $\mathbb{P}^2$.

Exploring this idea further, it is useful to view $\mathbb{P}^2$ as being covered by 3 copies of $\mathbb{C}^2$, though not disjointly. We consider three ways to embed $\mathbb{C}^2$ as an open subset of $\mathbb{P}^2$; these three embeddings send (a, b) to $(1 : a : b)$, $(a : 1 : b)$, and $(a : b : 1)$, respectively. We will denote the three images of $\mathbb{C}^2$ by U_1, U_2, and U_3, respectively. Notice that $\mathbb{P}^2 = U_1 \cup U_2 \cup U_3$, but that the union is not disjoint. In order to "see" a subset S of $\mathbb{P}^2$, it is often useful to look at $S \cap U_i$ for each i. For example, we compute $Y \cap U_2$.

Lemma 10.1.2. *We have*

$$Y \cap U_2 = \left\{ \left(\frac{x}{y} : 1 : \frac{1}{y} \right) : y^2 = x^3 + ax + b \text{ and } y \neq 0 \right\}.$$

In particular, Y is not closed in $\mathbb{P}^2$, since $(0 : 1 : 0)$ is a limit point of $Y \cap U_2$ that is not contained in $Y \cap U_2$.

Proof. The first equality follows easily from the definitions. To see that $(0 : 1 : 0)$ is a limit point, let $\alpha^{1/2}$ denote the positive square root of the positive real number α. Then $|x/y| = |x|/|x^3 + ax + b|^{1/2} \to 0$ and $|1/y| = 1/|x^3 + ax + b|^{1/2} \to 0$ as $|x| \to \infty$. $\square$

10.1.3 The Closure of $y^2 = x^3 + ax + b$ in $\mathbb{P}^2$

Proposition 10.1.3. *The closure of the graph Y of $y^2 = x^3 + ax + b$ in $\mathbb{P}^2$ is the graph X of $y^2 z = x^3 + axz^2 + bz^3$ in $\mathbb{P}^2$. We have*

$$\begin{aligned} X &= \{(x : y : z) \in \mathbb{P}^2 : y^2 z = x^3 + axz^2 + bz^3\} \\ &= Y \cup \{(0 : 1 : 0)\}. \end{aligned}$$

Proof. The inverse of image of X in $\mathbb{C}^3 \setminus \{0\}$ is the inverse image of 0 under the continuous map $\mathbb{C}^3 \setminus \{0\} \to \mathbb{C}$ defined by $(x, y, z) \mapsto y^2 z - (x^3 + axz^2 + bz^3)$, so X is closed. The difference $X \setminus Y$ is the set of points $(x : y : 0)$ that satisfy $y^2 z = x^3 + axz^2 + bz^3$, so $X \setminus Y = \{(0 : 1 : 0)\}$. Thus X is closed and $X = Y \cup \{(0 : 1 : 0)\}$. By Lemma 10.1.2, Y is not closed in $\mathbb{P}^2$, so X is the closure of Y in $\mathbb{P}^2$. $\square$

We will frequently refer to the point $(0 : 1 : 0)$ on X as the "point at infinity".

Definition 10.1.4 (Elliptic Curve). An *elliptic curve* E over the complex numbers $\mathbb{C}$ is a closed cubic curve in $\mathbb{P}^2$ defined by an equation

$$y^2 = x^3 + ax + b,$$

with $a, b \in \mathbb{C}$ and $\Delta = -16(4a^3 + 27b^2) \neq 0$. We let $E(\mathbb{C})$ denote the underlying set of points on E in $\mathbb{P}^2$.

If $4a^3 + 27b^2 = 0$ then the cubic $x^3 + ax + b$ has a repeated root α. Locally at $(\alpha, 0)$, Y is does not behave like an open subset of $\mathbb{C}$, which is why we exclude this case.

Remark 10.1.5. An ellipse is the graph of $ax^2 + by^2 = r$ with $a, b, r > 0$. Elliptic curves are not ellipses; they are called "elliptic" because they arise when studying arc lengths of ellipses (see Exercise 14). They haven't always been called elliptic curves; e.g., in the early 1960s Cassels called them "abelian varieties of dimension one" (see [Cas62]).

10.2 The Group Structure on an Elliptic Curve

Let E be an elliptic curve over $\mathbb{C}$. In this section we will see how to put an abelian group structure on the set $E(\mathbb{C})$ in a natural way. The point $\mathcal{O} = (0 : 1 : 0) \in E(\mathbb{C})$ "at infinity" will serve as the 0 element of the group, and if $P, Q, R \in E(\mathbb{C})$ then $P + Q + R = \mathcal{O}$ if and only if P, Q, and R lie on a common line. There is a sense, which we will not make precise, in which elliptic curves are the only curves that can be endowed in a "natural way" with a group structure.

10.2.1 Divisors

We begin by defining a huge group associated to E.

Definition 10.2.1 (Divisors). The group $\mathrm{Div}(E)$ of *divisors on* E is the free abelian group on the elements of $E(\mathbb{C})$. Thus $\mathrm{Div}(E)$ is the set of all finite formal linear combinations

$$n_1P_1 + n_2P_2 + \cdots + n_iP_i$$

with $n_i \in \mathbb{Z}$ and $P_i \in E(\mathbb{C})$.

Because $\mathrm{Div}(E)$ is a *free* abelian group, there are no relations among the points; thus, e.g., if P, Q, and R are in $E(\mathbb{C})$ then *by definition* we will never have $P + Q = R$ in $\mathrm{Div}(E)$.

Example 10.2.2. If E is defined by $y^2 = x(x-1)(x+1)$, then

$$2(0,0) - 3(1,0) + (-1,0) + (2,\sqrt{6})$$

is an element of $\mathrm{Div}(E)$.

Note that $\mathrm{Div}(E)$ is a huge uncountable group, and there is no natural way in which its elements are in bijection with $E(\mathbb{C})$; it is much too large. We will cut it down significantly, by considering its quotient by the subgroup of *principal divisors*. This is analogous to considering the quotient of the nonzero ideals of a quadratic field by the equivalence relation $\sim$ as in Section 9.5.3.

10.2.2 Rational Functions

Definition 10.2.3 (Rational Function on $\mathbb{P}^2$). A *rational function* on $\mathbb{P}^2$ is an element of the field $\mathbb{C}(x,y)$ of all quotients $p(x,y)/q(x,y)$ where $p(x,y)$ and $q(x,y)$ are arbitrary polynomials in two variables with $q \neq 0$.

Definition 10.2.4 (Homogenous Polynomial). A *homogeneous polynomial* in n-variables and of degree d is a polynomial $P(x_1,\ldots,x_n)$ such that for every constant λ,

$$P(\lambda x_1,\ldots,\lambda x_n) = \lambda^d P(x_1,\ldots,x_n).$$

Notice that $x^2 + y^2$ is homogenous polynomial, but $y^2 + x^3$ is not.

A rational function $f = p(x,y)/q(x,y) \in \mathbb{C}(x,y)$ determines an algebraic map $\mathbb{P}^2 \to \mathbb{P}^1$ as follows. Let $P(X,Y,Z) = Z^r p(X/Z, Y/Z)$ and $Q(X,Y,Z) = Z^s q(X/Z,Y/Z)$, where r and s are the degrees of p and q. If $s < r$, replace Q by $Z^{r-s}Q$ or if $r < s$ replace P by $Z^{s-r}P$, so that P and Q have the same degree. Then

$$(a\colon b\colon c) \mapsto (P(a,b,c) : Q(a,b,c))$$

is a well-defined algebraic map from $\mathbb{P}^2 \to \mathbb{P}^1$. It is *not* true that every map $\mathbb{P}^2 \to \mathbb{P}^1$ is induced by a rational function; for example, the constant function that sends every element of $\mathbb{P}^2$ to $(1 : 0)$ doesn't come from a rational function, since the rational function that induced it would have a denominator of 0.

Let E be the elliptic curve defined by $y^2 = x^3 + ax + b$.

Definition 10.2.5 (Rational Function on E). A *rational function* on E is an element of the field

$$K(E) = \mathbb{C}(x)[y]/(y^2 - (x^3 + ax + b)) = \mathbb{C}(x)(\sqrt{x^3 + ax + b}).$$

Thus $K(E)$ is the field generated by x and y where x is an indeterminate and y satisfies $y^2 = x^3 + ax + b$. Thus $K(E)$ is a quadratic field extension of $\mathbb{C}(x)$. Just as is the case for rational functions on $\mathbb{P}^2$, a rational function determines a map $E(\mathbb{C}) \to \mathbb{P}^1 = \mathbb{C} \cup \{\infty\}$ (but not conversely).

The analogue of the ring of integers of $K(E)$ is called the "affine coordinate ring" of E.

Definition 10.2.6 (Affine Coordinate Ring of E). The *affine coordinate ring* of E is the subgroup

$$A(E) = \mathbb{C}[x, y]/(y^2 - (x^3 + ax + b)) = \mathbb{C}[x][\sqrt{x^3 + ax + b}].$$

One can prove that $A(E)$ is integrally closed in $K(E)$ (see Exercise 4 for some examples).

Proposition 10.2.7. *There is a natural bijection between the maximal ideals of the ring $A(E)$ and the elements of $E(\mathbb{C}) \setminus \{\mathcal{O}\}$.*

Proof. If $(\alpha, \beta) \in \mathbb{C}^2$ is a point on $E(\mathbb{C})$ let $\mathfrak{m} = (x - \alpha, y - \beta)$ be the ideal in $A(E)$ generated by $x - \alpha$ and $y - \beta$. Since $\mathfrak{m}$ is the kernel of the homomorphism $A(E) \to \mathbb{C}$ sending x to α and y to β, we see that $\mathfrak{m}$ is a maximal ideal.

Conversely, suppose that $\mathfrak{m}$ is a maximal ideal of $A(E)$. The inverse image of $\mathfrak{m}$ in $\mathbb{C}[x]$ is a nonzero prime ideal of $\mathbb{C}[x]$ (it is a general fact that the inverse image of any prime ideal under any homomorphism is a prime ideal), so it is of the form $(x - \alpha)$ for some $\alpha \in \mathbb{C}$. Thus $A(E)/\mathfrak{m}$ is a quotient of

$$R = \mathbb{C}[y]/(y^2 - \alpha^3 - a\alpha - b).$$

Since $\mathbb{C}$ is algebraically closed, the maximal ideals of R correspond to the points on E with x-coordinate α. Thus $\mathfrak{m}$ corresponds to a point (α, β) on E. ☐

10.2.3 Principal Divisors

Let $K(E)^\times$ denote the group of nonzero elements of $K(E)$, and suppose $f = p(x, y)/q(x, y) \in K(E)^\times$, where $p(x, y), q(x, y) \in A(E)$. Let $P = (\alpha : \beta : 1) \in E(\mathbb{C})$ and let $\mathfrak{m} = (x - \alpha, y - \beta)$ be the corresponding maximal ideal of $A(E)$ as in Proposition 10.2.7. The order of vanishing of $p(x, y)$ at P is

$$\mathrm{ord}_P(p(x, y)) = \max\{n : p \in \mathfrak{m}^n\} \in \mathbb{Z},$$

and likewise the order of vanishing of $q(x, y)$ at P is

$$\mathrm{ord}_P(q(x, y)) = \max\{n : q \in \mathfrak{m}^n\} \in \mathbb{Z}.$$

The points P where $\mathrm{ord}_P(p(x, y)) > 0$, are the points where the graph of $p(x, y) = 0$ intersects E. Since $y^2 - (x^3 + ax + b)$ doesn't divide $p(x, y)$ (by assumption), there are only finitely many intersection points, so there are only finitely many points P where $\mathrm{ord}_P(p(x,y)) > 0$.

The *order* of the rational function f at P is

$$\mathrm{ord}_P(f) = \mathrm{ord}_P(p(x,y)) - \mathrm{ord}_P(q(x,y)) \in \mathbb{Z}.$$

Let $\mathcal{O} = (0:1:0)$ be the point at infinity on E. We define, in a seemingly totally ad hoc manner,

$$\mathrm{ord}_{\mathcal{O}}(f) = -\sum \mathrm{ord}_P(f) \in \mathbb{Z},$$

where the sum is over all $P = (\alpha : \beta : 1) \in E(\mathbb{C})$. This is the same value we would obtain if we were to define $A(E)$, etc., as above, but with U_3 replaced by U_2, but we will not prove this fact in this book (it is nontrivial).

Definition 10.2.8 (Divisor of a Function). Let $f \in K(E)^\times$. The *principal divisor* associated to f is

$$(f) = \sum_{\text{all } P \in E(\mathbb{C})} \mathrm{ord}_P(f) \cdot P \in \mathrm{Div}(E).$$

The map $K(E)^\times \to \mathrm{Div}(E)$ is a group homomorphism; this is just the assertion that the order of vanishing at a point P of the product of two functions on E is the sum of their orders of vanishing at P, a fact we will not prove in this book.

10.2.4 Picard's Group and the Group Law

Let $\mathrm{Prin}(E)$ be the subgroup of $\mathrm{Div}(E)$ of principal divisors.

Definition 10.2.9 (The Picard Group). The *Picard group* of E is

$$\mathrm{Pic}(E) = \mathrm{Div}(E)/\mathrm{Prin}(E).$$

Alternatively, $\mathrm{Pic}(E)$ is the set of equivalence classes of elements of $\mathrm{Div}(E)$ with respect to the equivalence relation $\sim$ in which $D_1 \sim D_2$ if and only if there is a rational function $f \in K(E)^\times$ such that $D_1 - D_2 = (f)$.

The Picard group is much smaller than $\mathrm{Div}(E)$ and has a more interesting structure. It is still too big though.

Definition 10.2.10 (Degree). The *degree* of a divisor $\sum n_i P_i \in \mathrm{Div}(E)$ is $\sum n_i \in \mathbb{Z}$. Suppose f is a nonzero rational function on E with divisor $(f) = \sum n_i P_i$. Then the *degree* of f is the sum of the n_i such that n_i is positive.

Notice that the map $\mathrm{Div}(E) \to \mathbb{Z}$ is a group homomorphism. Let $\mathrm{Div}^0(E)$ denote the subgroup of divisors of degree 0. Because of how we defined $\mathrm{ord}_{\mathcal{O}}(f)$ for $\mathcal{O} = (0:1:0)$, it is trivially true that $\mathrm{Prin}(E) \subset \mathrm{Div}^0(E)$. Let

$$\mathrm{Pic}^0(E) = \mathrm{Div}^0(E)/\mathrm{Prin}(E).$$

Lemma 10.2.11. *There are no rational functions of degree* 1 *on an elliptic curve.*

Proof. A rational function of degree 1 would define an isomorphism between E and $\mathbb{P}^1$, which is impossible because E is a torus and $\mathbb{P}^1$ is a sphere.[2] □ 2

Theorem 10.2.12. *The map $\Phi : E(\mathbb{C}) \to \mathrm{Pic}^0(E)$ that associates to a point $P \in E(C)$ the class of the degree 0 divisor $P - \mathcal{O}$ is a bijection. Since $\mathrm{Pic}^0(E)$ is a group, this bijection induces a group structure on $E(\mathbb{C})$. In this group, if P, Q, and R are colinear points on E, then $P + Q + R = 0$.*

Proof. First we show that Φ is injective. If $\Phi(P) = \Phi(Q)$ with $P \neq Q$, then $P - \mathcal{O} \sim Q - \mathcal{O}$. Thus $P \sim Q$, so there is a rational function f on E of degree 1, which contradicts Lemma 10.2.11. Thus Φ is injective.

To show that Φ is surjective, we must show that every element of $\mathrm{Div}^0(E)$ is equivalent to an element of the form $P - \mathcal{O}$ for some $P \in E(\mathbb{C})$. Suppose $\sum n_i P_i$ is an element of $\mathrm{Div}^0(E)$. Then $\sum n_i P_i = \sum n_i (P_i - \mathcal{O})$ since $\sum n_i = 0$. By induction it thus suffices to show that $(P - \mathcal{O}) \pm (Q - \mathcal{O}) \sim R - \mathcal{O}$ for some R. We do this using rational functions of the form $f = cx + dy + e$. Because E is defined by a cubic equation $y^2 = x^3 + ax + b$, the divisor of f is

$$(f) = P + Q + R - 3\mathcal{O},$$

where P, Q, and R are the three points of intersection of the line $f = 0$ with E, counted with multiplicity. Thus

$$(P - \mathcal{O}) + (Q - \mathcal{O}) + (R - \mathcal{O}) \sim 0.$$

If $R = (x, y)$, let $\tilde{R} = (x, -y)$. Then using a vertical line we see that $R + \tilde{R} \sim 2\mathcal{O}$, so

$$(R - \mathcal{O}) + (\tilde{R} - \mathcal{O}) \sim 0.$$

Thus $(P - \mathcal{O}) + (Q - \mathcal{O}) \sim (\tilde{R} - \mathcal{O})$. Likewise, $(P - \mathcal{O}) - (Q - \mathcal{O}) \sim (P - \mathcal{O}) + (\tilde{Q} - \mathcal{O})$, so $(P - \mathcal{O}) - (Q - \mathcal{O})$ is equivalent to a divisor of the desired form. This completes the proof. □

Remark 10.2.13. If E is replaced by a plane curve X of higher degree (without singularities), then everything that we stated about divisors is true except the theorem, which is false. Instead we have only an injective map $X \hookrightarrow \mathrm{Pic}^0(X)$. The group $\mathrm{Pic}^0(X)$ is called the *Jacobian* of X and has additional geometric structure (e.g., its elements are in natural bijection with the points on an algebraic variety of dimension $(d-1)(d-2)/2$, where d is the degree of X). Thus, though $X(\mathbb{C})$ doesn't have a natural group structure, it embeds in a geometric object which does.

Just as for binary quadratic forms, the group structure is induced by multiplication of ideal classes. Let $\mathcal{I}$ denote the set of nonzero ideals of the affine coordinate ring $A(E)$ of E. One can prove that every element of $\mathcal{I}$ is a product of maximal ideals of $A(E)$. By Proposition 10.2.7, these maximal ideals are in bijection with the points $E(\mathbb{C})$. Define an equivalence relation $\sim$ on $\mathcal{I}$ by $I \sim J$ if there are nonzero $f, g \in A(E)$ such that $(f)I = (g)J$, and let $\mathrm{Cl}(E)$ denote the group of equivalence classes of nonzero ideals

[2] This proof is incomplete, given the theory developed thus far in this course.

under multiplication. Then the map $\mathrm{Cl}(E) \to \mathrm{Pic}^0(E)$ which sends the class of the maximal ideal $\mathfrak{m}$ corresponding to a point P to the class of the divisor $P - \mathcal{O}$ is an isomorphism.

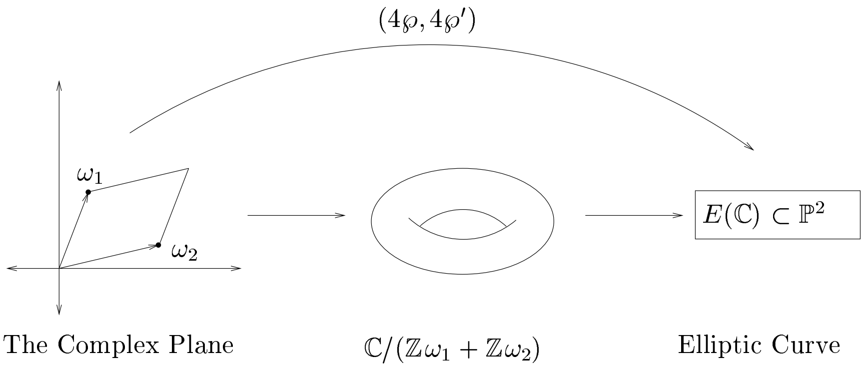


FIGURE 10.2. The Weierstrass $\wp$ Function and Elliptic Curves

10.2.5 Analytic Description of the Group Law

An alternative approach to the group law is via the Weierstrass $\wp$ function from complex analysis (see, e.g., [Ahl78,]). Let a and b be complex number with $4a^3 + 27b^2 \neq 0$. The Weierstrass $\wp$ function associated to a and b is a function $\wp : \mathbb{C} \to \mathbb{C} \cup \{\infty\}$ whose set of poles (points that map to ∞) are of the form $\mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$, where $\omega_1, \omega_2 \in \mathbb{C}$ have the property that $\mathbb{R}\omega_1 + \mathbb{R}\omega_2 = \mathbb{C}$. Moreover, $\wp$ is periodic with periods ω_1 and ω_2, in the sense that $\wp(z + n_1\omega_1 + n_2\omega_2) = \wp(z)$ for all $z \in \mathbb{C}$.

The connection with elliptic curves via $\wp$ is illustrated in Figure 10.2. If $x = 4\wp$ and $y = 4\wp'$, then $y^2 = x^3 + ax + b$, and $z \mapsto (4\wp(z), 4\wp'(z))$ extends to a complex analytic bijection

$$f : \mathbb{C}/(\mathbb{Z}\omega_1 + \mathbb{Z}\omega_2) \to E(\mathbb{C}).$$

Since $\mathbb{C}/(\mathbb{Z}\omega_1 + \mathbb{Z}\omega_2)$ is an abelian group, f induces a group structure on E, and one can show that this group structure is the same as the one obtained above using divisors. Also note that the quotient $\mathbb{C}/(\mathbb{Z}\omega_1 + \mathbb{Z}\omega_2)$ is topologically homeomorphic to a torus.

10.2.6 Geometric Description of the Group Law

Suppose $a, b \in \mathbb{R}$ and consider the elliptic curve defined by $y^2 = x^3 + ax + b$. Since $a, b \in \mathbb{R}$ it makes sense to draw a graph of the subset

$$E(\mathbb{R}) = \{(x, y) \in \mathbb{R} \times \mathbb{R} : y^2 = x^3 + ax + b\} \subset E(\mathbb{C}).$$

Qualitatively, there are two possibilities for this picture, depending on whether $x^3 + ax + b$ has 1 or 3 real roots. If $x^3 + ax + b$ has 1 real root, then $E(\mathbb{R})$ looks like a bow, and if the cubic has 3 real roots, then $E(\mathbb{R})$ looks like the union of an egg and a bow. See Figure 10.3 for graphs of $y^2 = x^3 + 1$ (one real root) and $y^2 = x^3 - 5x + 4$ (three real roots).

Geometrically, the group law on $E(\mathbb{R})$ is defined as illustrated in Figure 10.4. To compute $P + Q$ draw the line L detetermined by $P + Q$ (if $P = Q$ let L be the tangent line to P). Let $R = (x, y)$ be the third point of intersection. Then the sum of $P + Q$ is $(x, -y)$. It is easy to see that the binary operation $E(\mathbb{R}) \times E(\mathbb{R}) \to E(\mathbb{R})$ determined by this process satisfies

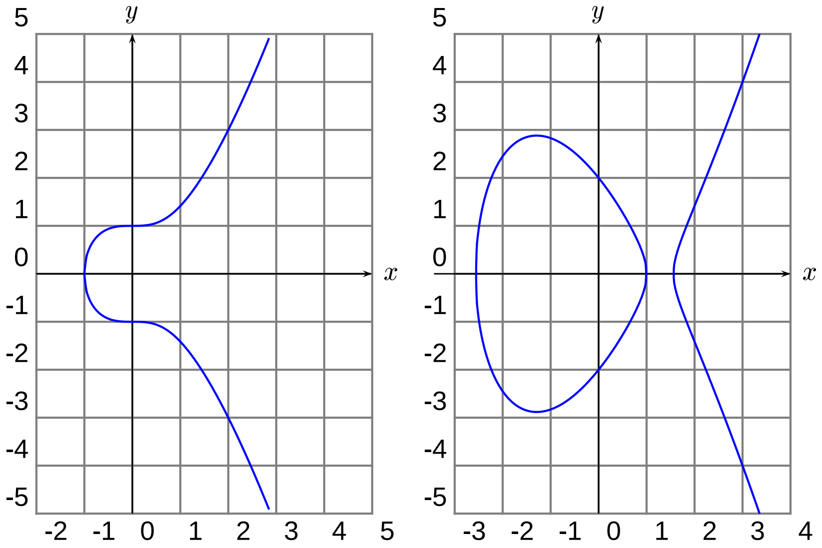


FIGURE 10.3. Real Points on $y^2 = x^3 + 1$ and $y^2 = x^3 - 5x + 4$

all the axioms of a group, except the associative law, which is much more tedious. (Note: One can rephrase all this geometry in a purely algebraic way, and then there is no need to restrict to real points.)

10.2.7 An Example

Let E be the elliptic curve defined by $y^2 = x^3 - 5x + 4$. Then $P = (0, 2)$, $Q = (1, 0)$, $R = (3, 4)$ and $R' = (3, -4)$ are elements of $E(\mathbb{C})$ and, as illustrated in Figure 10.4, $P + Q = R$. We verify this from the point of view of divisors and the Weierstrass $\wp$ function.

From the point of view of divisors, $P + Q = R$ is the assertion that

$$P - \mathcal{O} + Q - \mathcal{O} \sim R - \mathcal{O}.$$

To verify this, we exhibit a rational function f such that

$$(f) = P - \mathcal{O} + Q - \mathcal{O} - (R - \mathcal{O}) = P + Q - R - \mathcal{O},$$

i.e., so that f has simple zeros at P and Q and simples poles at R and $\mathcal{O}$. Let $f = \frac{2x+y-2}{x-3}$. Then

$$\begin{aligned}(2x + y - 2) &= P + Q + R' - 3\mathcal{O}\\ (x - 3) &= R + R' - 2\mathcal{O},\end{aligned}$$

so

$$(f) = P + Q + R' - 3\mathcal{O} - (R + R' - 2\mathcal{O}) = P + Q - R - \mathcal{O}$$

as required.

Let $\wp$ be the Weierstrass function associated to $y^2 = x^3 - 5x + 4$. The poles of $\wp$ are the elements of $\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$ where

$$\omega_1 = 2.3970980311644782804\ldots, \quad \omega_2 = i1.6043106621845101475\ldots.$$

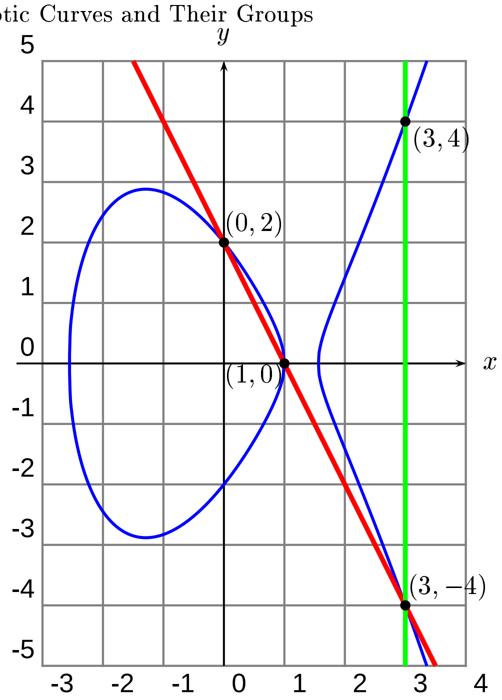


FIGURE 10.4. The Group Law: $(1,0)+(0,2)=(3,4)$ on $y^2=x^3-5x+4$

Under the map $z \mapsto (4\wp(z), 4\wp'(z))$ we have

$$z_P = 0.58916472693707629\ldots + i0.8021553310922550\ldots \mapsto P$$
$$z_Q = 1.19854901558223914\ldots + i0.8021553310922588\ldots \mapsto Q$$
$$z_R = 1.78771374251931543\ldots \mapsto R$$

We have $z_P + z_Q = z_R + \omega_2$, so $z_P + z_Q = z_R \pmod{\Lambda}$, as expected.

10.2.8 Equations for the Group Law

In this section we give a description of the group law in terms of equations. Suppose that $P=(x_1,y_1)$ and $Q=(x_2,y_2)$ are nonzero points on $y^2 = x^3+ax+b$. If $P \neq \pm Q$, let $\lambda = (y_1-y_2)/(x_1-x_2)$ and $\nu = y_1 - \lambda x_1$. Then $P+Q=(x_3,y_3)$ where

$$x_3 = \lambda^2 - x_1 - x_2 \quad \text{and} \quad y_3 = -\lambda x_3 - \nu.$$

If $P=-Q$ (i.e., $x_1=x_2$ and $y_1=-y_2$), then $P+Q=0$. If $P=Q$ (but $P\neq -Q$) then

$$x_3 = \frac{(x_1^2-a)^2 - 8bx_1}{4y_1^2},$$
$$y_3 = \frac{(3x_1^2+a)(x_1-x_3) - 2y_1^2}{2y_1}.$$

Note that in case $P \neq Q$, the group law equations do not involve a and b! However, in this case P and Q completely determine a and b (see Exercise 10).

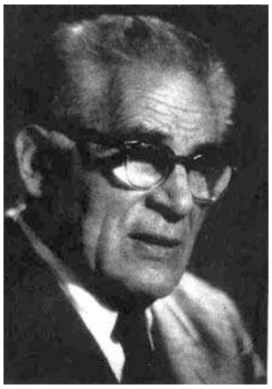


FIGURE 10.5. Louis J. Mordell

10.3 Rational Points

Choose $a, b \in \mathbb{C}$ and consider the abelian group $E(\mathbb{C})$ associated to $y^2 = x^3 + ax + b$. As described in Section 10.2.5, $E(\mathbb{C})$ is isomorphic to $\mathbb{C}/(\mathbb{Z}\omega_1 + \mathbb{Z}\omega_2)$ for complex numbers ω_1 and ω_2. Viewing $\mathbb{C}$ as a two-dimensional real vector space with basis ω_1 and ω_2, we see that

$$E(\mathbb{C}) \cong \mathbb{C}/(\mathbb{Z}\omega_1 + \mathbb{Z}\omega_2) \cong (\mathbb{R}/\mathbb{Z}) \oplus (\mathbb{R}/\mathbb{Z}).$$

Thus, as an abstract abelian group, $E(\mathbb{C})$ does not depend on the elliptic curve E!

A *number field* is a field K that contains $\mathbb{Q}$ and is finite dimensional when viewed as a $\mathbb{Q}$-vector space. Think of K as being obtained from $\mathbb{Q}$ by "adjoining" to $\mathbb{Q}$ a root of a polynomial with coefficients in $\mathbb{Q}$. For example, $K = \mathbb{Q}$ is a number field, and we studied number fields of dimension 2 over $\mathbb{Q}$ in Chapter 9.

When $a, b \in K$ we say that the elliptic curve E associated to $y^2 = x^3 + ax + b$ is *defined over* K. Then the subset

$$E(K) = \{(x, y) \in K \times K : y^2 = x^3 + ax + b\} \cup \{\mathcal{O}\} \subset E(\mathbb{C})$$

is a subgroup, the group of points on E *rational over* K.

The groups $E(K)$ are much more interesting than $E(\mathbb{C})$. For example, if E is the elliptic curve defined by $y^2 = x^3 - 5x + 4$ from Section 10.2.7, then

$$E(\mathbb{Q}) \cong \mathbb{Z} \times (\mathbb{Z}/2),$$

where a generator for the $\mathbb{Z}$-factor is the point $(0, -2)$ and the generator for the $\mathbb{Z}/2$ factor is $(1, 0)$.

Let E be an elliptic curve defined over a number field K.

Theorem 10.3.1 (Louis Mordell). *The group $E(K)$ is finitely generated. That is, there are points $P_1, \ldots, P_s \in E(K)$ such that every element of $E(K)$ is of the form $n_1P_1 + \cdots + n_sP_s$ for integers $n_1, \ldots n_s \in \mathbb{Z}$.*

Because of this theorem, the group $E(K)$ is often called the *Mordell-Weil group* of E over K. (Andre Weil generalized Mordell's theorem to abelian varieties, which are higher-dimensional analogues of elliptic curves.)

TABLE 10.1. Exhibiting Every Possible Torsion Subgroup Over $\mathbb{Q}$

Curve	$E(\mathbb{Q})_{\mathrm{tor}}$
$y^2 = x^3 - 2$	$\{0\}$
$y^2 = x^3 + 8$	$\mathbb{Z}/2$
$y^2 = x^3 + 4$	$\mathbb{Z}/3$
$y^2 = x^3 + 4x$	$\mathbb{Z}/4$
$y^2 - y = x^3 - x^2$	$\mathbb{Z}/5$
$y^2 = x^3 + 1$	$\mathbb{Z}/6$
$y^2 = x^3 - 43x + 166$	$\mathbb{Z}/7$
$y^2 + 7xy = x^3 + 16x$	$\mathbb{Z}/8$
$y^2 + xy + y = x^3 - x^2 - 14x + 29$	$\mathbb{Z}/9$
$y^2 + xy = x^3 - 45x + 81$	$\mathbb{Z}/10$
$y^2 + 43xy - 210y = x^3 - 210x^2$	$\mathbb{Z}/12$
$y^2 = x^3 - 4x$	$\mathbb{Z}/2 \times \mathbb{Z}/2$
$y^2 = x^3 + 2x^2 - 3x$	$\mathbb{Z}/4 \times \mathbb{Z}/2$
$y^2 + 5xy - 6y = x^3 - 3x^2$	$\mathbb{Z}/6 \times \mathbb{Z}/2$
$y^2 + 17xy - 120y = x^3 - 60x^2$	$\mathbb{Z}/8 \times \mathbb{Z}/2$

Mordell's theorem implies that it makes sense to ask whether or not we can compute $E(K)$, where by "compute" we mean find a finite set $P_1, \ldots, P_s$ of points on E that generate $E(K)$. There is a systematic theory that addresses the question of how to compute $E(K)$ (see, e.g., [Sil86]); in practice this theory often produces answers, but we don't know that it always will.

Conjecture 10.3.2. *There is an algorithm that given an elliptic curve E over a number field K outputs a finite list of generates for $E(K)$.*

Note that this is not a conjecture about computational complexity. The conjecture is that there is an algorithm to compute $E(K)$, not that $E(K)$ can be computed quickly.

10.3.1 The Torsion Subgroup and the Rank

The set of elements of $E(K)$ of finite order is a subgroup of $E(K)$ which we denote by $E(K)_{\mathrm{tor}}$. For example, if E is defined by $y^2 = x^3 - 5x + 4$, then

$$E(\mathbb{Q})_{\mathrm{tor}} = \{\mathcal{O}, (1,0)\} \cong \mathbb{Z}/2.$$

Theorem 10.3.3 (Mazur, 1976). *Let E be an elliptic curve over $\mathbb{Q}$. Then $E(\mathbb{Q})_{\mathrm{tor}}$ is isomorphic to one of the following 15 groups:*

$$\begin{aligned} \mathbb{Z}/n &\qquad \text{for } n \leq 10 \text{ or } n = 12, \\ \mathbb{Z}/2 \times \mathbb{Z}/2n &\qquad \text{for } n \leq 4. \end{aligned}$$

Table 10.1 lists elliptic curves with each possible torsion subgroup.

The quotient $E(K)/E(K)_{\mathrm{tor}}$ is a finitely generated free abelian group, so it is isomorphism to $\mathbb{Z}^r$ for some integer r, called the *rank* of $E(K)$.

Conjecture 10.3.4. *There are elliptic curves over $\mathbb{Q}$ of arbitrarily large rank.*

Probably nobody has a clue as to how to prove Conjecture 10.3.4. The "world record" is a curve of rank ≥ 24. It was discovered in January 2000 by Roland Martin and William McMillen of the National Security Agency. They weren't allowed to tell how they found the curve, and for several months they could only announce that they found a curve of rank ≥ 24, but they not release the curve to the public. Here it is:

Proposition 10.3.5. *The elliptic curve*

$$y^2+xy+y = x^3 - 120039822036992245303534619191166796374x$$
$$+ 504224992484910670010801799168082726759443756222911415116$$

over $\mathbb{Q}$ has rank at least 24. The following points $P_1, \ldots, P_{24}$ are independent points on the curve:

$$
\begin{aligned}
P_1 &= (2005024558054813068, -1648037158834308510823488252)\\
P_2 &= (-4690836759490453344, -31049883525785801514744524804)\\
P_3 &= (4700156326649806635, -6622116250158424945781859743)\\
P_4 &= (6785546256295273860, -1456180928830978521107520473)\\
P_5 &= (6823803569166584943, -1685950735477175947351774817)\\
P_6 &= (7788809602110240789, -6462981622972389783453855713)\\
P_7 &= (27385442304350994620556, 4531892554281655472841805111276996)\\
P_8 &= (54284682060285253719/4, -296608788157989016192182090427/8)\\
P_9 &= (-94200235260395075139/25, -3756324603619419619213452459781/125)\\
P_{10} &= (-3463661055331841724647/576,\\
&\qquad - 439033541391867690041114047287793/13824)\\
P_{11} &= (-6684065934033506970637/676,\\
&\qquad - 473072253066190669804172657192457/17576)\\
P_{12} &= (-956077386192640344198/2209,\\
&\qquad - 2448326762443096987265907469107661/103823)\\
P_{13} &= (-27067471797013364392578/2809,\\
&\qquad - 4120976168445115434193886851218259/148877)\\
P_{14} &= (-25538866857137199063309/3721,\\
&\qquad - 7194962289937471269967128729589169/226981)\\
P_{15} &= (-1026325011760259051894331/108241,\\
&\qquad - 1000895294067489857736110963003267773/35611289)\\
P_{16} &= (9351361230729481250627334/1366561,\\
&\qquad - 2869749605748635777475372339306204832/1597509809)\\
P_{17} &= (10100878635879432897339615/1423249,\\
&\qquad - 5304965776276966451066900941489387801/1697936057)\\
P_{18} &= (11499655868211022625340735/17522596,\\
&\qquad - 1513435763341541188265230241426826478043/73349586856)\\
P_{19} &= (110352253665081002517811734/21353641,\\
&\qquad - 461706833308406671405570254542647784288/98675175061)\\
P_{20} &= (414280096426033094143668538257/285204544,\\
&\qquad 266642138924791310663963499787603019833872421/4816534339072)\\
P_{21} &= (36101712290699828042930087436/4098432361,\\
&\qquad - 2995258855766764520463389153587111670142292/262377541318859)\\
P_{22} &= (45442463408503524215460183165/5424617104,\\
&\qquad - 3716041581470144108721590695554670156388869/399533898943808)\\
P_{23} &= (983886013344700707678587482584/141566320009,\\
&\qquad - 126615818387717930449161625960397605741940953/53264752602346277)\\
P_{24} &= (1124614335716851053281176544216033/152487126016,\\
&\qquad - 37714203831317877163580088877209977295481388540127/59545612760743936)
\end{aligned}
$$

Proof. See [MW00]. □

Given rational numbers a and b (with $4a^3 + 27b^2 \neq 0$) we have an associated group $E(\mathbb{Q})$, and a rank r. Even after several decades, mathematicians have given no examples of a, b for which they could show that $r > 24$, yet they still conjecture that r can be arbitrary large.

And they are probably right. Where does this intuition come from?

EXERCISES

10.1 We call a line in $\mathbb{C}^2$ *rational* if it is the set of zeros of an equation $ax + by + c = 0$ with $a, b, c \in \mathbb{Q}$.

(a) Suppose P and Q are distinct elements of $\mathbb{Q}^2$. Prove that the unique line in $\mathbb{C}^2$ that contains P and Q is rational.

(b) Suppose that L_1 and L_2 are distinct rational lines in $\mathbb{C}^2$ that intersect. Prove that their intersection is a rational point.

10.2 Let $Y \subset \mathbb{C}^2$ be the set of complex solutions (x, y) to the equation $y^2 = x^5 + 1$. Find (with proof!) the closure of Y in $\mathbb{P}^2$.

10.3 Let E be the elliptic curve defined by $y^2 = x^3 + 1$. Find the divisor associated to the rational function $(x+1)/(y-1)$.

10.4 Let x and y be indeterminates.

(a) Prove that $\mathbb{C}[x,y]/(y^2-(x^3+1))$ is integrally closed in $\mathbb{C}(x)[y]/(y^2-(x^3+1))$. That is, if $f(x), g(x) \in \mathbb{C}(x)$ are rational functions in x and $f(x) + yg(x)$ satisfies a monic polynomial with coefficients in $\mathbb{C}(x)$, then $f(x)$ and $g(x)$ are polynomials.

(b) Prove that $\mathbb{C}[x,y]/(y^2-x^3)$ is *not* integrally closed in $\mathbb{C}(x)[y]/(y^2-x^3)$. (Hint: Consider $t = y/x$.)

10.5 Let E be the elliptic curve defined by $y^2 = x^3 + x + 1$. Consider the points $P = (72\colon -611\colon 1)$, $Q = (1/4\colon -9/8\colon 1)$, and $R = (1\colon \sqrt{3}\colon 1)$ on E.

(a) Compute the sum of P and Q on E.

(b) Find nonzero integers n and m such that $nP = mQ$.

(c) Compute $R + R$.

(d) Is there any integer n such that $nR = P$? (Hint: Keep in mind the automorphism $\sqrt{3} \mapsto -\sqrt{3}$ of $\mathbb{Q}(\sqrt{3})$.)

10.6 Draw a graph of the set $E(\mathbb{R})$ of real points on each of the following elliptic curves:

(a) $y^2 = x^3 - 1296x + 11664$,

(b) $y^2 + y = x^3 - x$,

(c) $y^2 + y = x^3 - x^2 - 10x - 20$.

10.7 A rational solution to the equation $y^2 - x^3 = -2$ is $(3, 5)$. Find a rational solution with $x \neq 3$ by drawing the tangent line to $(3, 5)$ and computing the third point of intersection.

10.8 Suppose $y^2 = x^3 + a_1x + b_1$ and $y^2 = x^3 + a_2x + b_2$ define two elliptic curves E_1 and E_2 over $\mathbb{C}$. Suppose that there are points $P, Q \in E_1(\mathbb{C}) \cap E_2(\mathbb{C})$ such that $P \neq \pm Q$. Prove that $a_1 = a_2$ and $b_1 = b_2$. (Hint: Characterize the set of common solutions to the two equations $y^2 = x^3 + a_1x + b_1$ and $y^2 = x^3 + a_2x + b_2$.)

10.9 Consider the elliptic curve $y^2 + xy + y = x^3$ over $\mathbb{Q}$. Find a linear change of variables that transforms this curve into a curve of the form $Y^2 = X^3 + aX + b$ for rational numbers a and b.

10.10 Let X be a nonempty set. Show that there exists a binary operation $X \times X \to X$ that endows X with the structure of group, as follows:

(a) If X is finite, there is a bijection between X and a cyclic group.

(b) If X is any infinite set then a nontrivial theorem in set theory, which is proved using Zorn's lemma, is that there is a bijection between X and $X \times X$ (for a proof, see [Hal60, §24]). Another theorem is that if there is an injection $X \hookrightarrow Y$ and an injection $Y \hookrightarrow X$, then there is a bijection $X \to Y$. Assuming these two facts, prove that the there is a bijection between X and the set of finite sequences of elements of X. (Hint: Consider the countable disjoint union

$$W = X \cup (X \times X) \cup (X \times X \times X) \cup \cdots.$$

Prove that there is a bijection between W and X, by showing that there is a bijection between W and $X \times \mathbb{Z}$, and that there is a bijection between $X \times \mathbb{Z}$ and X.)

(c) If X is infinite, let A be the free abelian group on the elements of X (just like $\mathrm{Div}(E)$ is the free abelian group on the points of E). Using the ideas from part (ii), prove that there is a bijection between X and A, so that X can be endowed with an abelian group structure.

10.11 Let E be the elliptic curve over the finite field $K = \mathbb{Z}/5\mathbb{Z}$ defined by the equation

$$y^2 = x^3 + x + 1.$$

(a) List all 9 elements of $E(K)$.

(b) What is the structure of the group $E(K)$, as a product of cyclic groups?

10.12 Let E be an elliptic curve over $\mathbb{R}$. Define a binary operation $\boxplus$ on $E(\mathbb{R})$ as follows:

$$P \boxplus Q = -(P + Q).$$

Thus the $\boxplus$ of P and Q is the third point of intersection of the line through P and Q with E.

(a) Lists the axiom(s) of a group that fail for $E(\mathbb{R})$ equipped with this binary operation. (The group axioms are "identity", "inverses", and "associativity".)

(b) Under what conditions on $E(\mathbb{Q})$ does this binary operation define a group structure on $E(\mathbb{Q})$? (E.g., when $E(\mathbb{Q}) = \{\mathcal{O}\}$ this binary operation does define a group.)

10.13 Let $g(t)$ be a quartic polynomial with distinct (complex) roots, and let α be a root of $g(t)$. Let $\beta \neq 0$ be any number.

(a) Prove that the equations

$$x = \frac{\beta}{t-\alpha}, \qquad y = x^2 u = \frac{\beta^2 u}{(t-\alpha)^2}$$

give an "algebraic transformation" between the curve $u^2 = g(t)$ and the curve $y^2 = f(x)$, where $f(x)$ is the cubic polynomial

$$f(x) = g'(\alpha)\beta x^3 + \frac{1}{2}g''(\alpha)\beta^2 x^2 + \frac{1}{6}g'''(\alpha)\beta^3 x + \frac{1}{24}g''''(\alpha)\beta^4.$$

(b) Prove that if g has distinct (complex) roots, then f also has distinct roots, and so $u^2 = g(t)$ is an elliptic curve.

10.14 In this problem you will finally find out exactly why elliptic curves are called "elliptic curves"! Let $0 < \beta \le \alpha$, and let C be the ellipse

$$\frac{x^2}{\alpha^2} + \frac{y^2}{\beta^2} = 1.$$

(a) Prove that the arc length of C is given by the integral

$$4\alpha \int_0^{\pi/2} \sqrt{1 - k^2 \sin^2\theta}\, d\theta$$

for an appropriate choice of constant k depending on α and β.

(b) Check your value for k in (i) by verifying that when $\alpha = \beta$, the integral yields the correct value for the arc length of a circle.

(c) Prove that the integral in (i) is also equal to

$$4\alpha \int_0^1 \sqrt{\frac{1-k^2t^2}{1-t^2}}\,dt = 4\alpha \int_0^1 \frac{1-k^2t^2}{\sqrt{(1-t^2)(1-k^2t^2)}}\,dt.$$

(d) Prove that if the ellipse E is not a circle, then the equation

$$u^2 = (1-t^2)(1-k^2t^2)$$

defines an elliptic curve (cf. the previous exercise). Hence the problem of determining the arc length of an ellipse comes down to evaluating the integral

$$\int_0^1 \frac{1-k^2t^2}{u}\,dt$$

on the "elliptic" curve $u^2 = (1-t^2)(1-k^2t^2)$.

10.15 Suppose that $P = (x, y)$ is a point on the cubic curve

$$y^2 = x^3 + ax + b.$$

(a) Verify that the x coordinate of the point $2P$ is given by the duplication formula

$$x(2P) = \frac{x^4 - 2ax^2 - 8bx + a^2}{4y^2}.$$

(b) Derive a similar formula for the y coordinate of $2P$ in terms of x and y.

(c) Find a polynomial in x whose roots are the x-coordinates of the points $P = (x, y)$ satisfying $3P = \mathcal{O}$. (Hint: The relation $3P = \mathcal{O}$ can also be written $2P = -P$.)

(d) For the particular curve $y^2 = x^3 + 1$, solve the equation in (iii) to find all of the points satisfying $3P = \mathcal{O}$. Note that you will have to use complex numbers.

10.16 Let Φ be the set of the 15 possible groups of the form $E(\mathbb{Q})_{\mathrm{tor}}$ for E an elliptic curve over $\mathbb{Q}$ (see Lecture 27). For each group $G \in \Phi$, if possible, find a finite field $k = \mathbb{Z}/p\mathbb{Z}$ and an elliptic curve E over k such that $E(k) \approx G$. (Hint: It is a fact that $|p + 1 - \#E(\mathbb{Z}/p\mathbb{Z}))| \leq 2\sqrt{p}$, so you only have to try finitely many p to show that a group G does not occur as the group of points on an elliptic curve over a finite field.)

10.17 Let E be the elliptic curve defined by the equation $y^2 = x^3 + 1$.

(a) For each prime p with $5 \leq p < 30$, describe the group of points on this curve having coordinates in the finite field $\mathbb{Z}/p\mathbb{Z}$. (You can just give the order of each group.)

(b) For each prime in (i), let N_p be the number of points in the group. (Don't forget the point infinity.) For the set of primes satisfying $p \equiv 2 \pmod 3$, can you see a pattern for the values of N_p? Make a general conjecture for the value of N_p when $p \equiv 2 \pmod 3$.

(c) Prove your conjecture.

10.18 Let E be an elliptic curve over the real numbers $\mathbb{R}$. Prove that $E(\mathbb{R})$ is not a finitely generated abelian group.

11

Algorithmic Applications of Elliptic Curves

[1] We assume the reader has read Chapter 10 about elliptic curves. 1

11.1 Elliptic Curves Over $\mathbb{Z}/p$

All of the applications of elliptic curves that we will discuss in this chapter involve elliptic curves over finite fields (or rings).

Let p be a prime and let $\mathbb{F}_p = \mathbb{Z}/p$ be the field with p elements. There is an analogue $\mathbb{P}^2_{\mathbb{F}_p}$ of the projective plane from Section 10.1.2 over the field $\mathbb{F}_p$. The set of points $\mathbb{P}^2(\mathbb{F}_p)$ of $\mathbb{P}^2_{\mathbb{F}_p}$ rational over $\mathbb{F}_p$ is the set of triples $(a : b : c)$ with $a, b, c \in \mathbb{F}_p$ not all zero modulo the equivalence relation in which $(\lambda a : \lambda b : \lambda c) = (a : b : c)$ for any nonzero $\lambda \in \mathbb{F}_p$.

Definition 11.1.1 (Elliptic Curve Over $\mathbb{F}_p$). An elliptic curve over $\mathbb{F}_p$ is a (closed) curve in $\mathbb{P}^2_{\mathbb{F}_p}$ defined by an equation of the form $y^2 = x^3 + ax + b$ such that $a, b \in \mathbb{F}_p$ and $4a^3 + 27b^2 \neq 0$.

Remark 11.1.2. It is more natural to define an elliptic curve to be a "nonsingular" plane cubic curve in $\mathbb{P}^2_{\mathbb{F}_p}$ equipped with a distinguished $\mathbb{F}_p$-rational point. When $p \geq 5$, every such curve can be transformed by a change of variables into a curve of the form $y^2 = x^3 + ax + b$. When $p = 2, 3$, this is not the case. For the applications in this chapter, we may assume that $p \geq 5$.

The set of points on an elliptic curve over $\mathbb{F}_p$ is

$$E(\mathbb{F}_p) = \{(x, y) : y^2 = x^3 + ax + b\} \cup \{(0 : 1 : 0)\},$$

[1] Write introduction to the chapter.

where, as usual, we write (a, b) for $(a : b : 1) \in \mathbb{P}^2(\mathbb{F}_p)$. Just as was the case for $E(\mathbb{C})$, the set $E(\mathbb{F}_p)$ is equipped with a natural group structure.

11.1.1 The Possibilities for $E(\mathbb{F}_p)$

In contrast to the situation with $E(\mathbb{Q})$ (see Section 10.3), the possibilities for the group $E(\mathbb{F}_p)$ are well understood.

Theorem 11.1.3. *The finite abelian group $E(\mathbb{F}_p)$ is either cyclic or a product of two cyclic groups.*

Proof. We only sketch the proof. Since $E(\mathbb{F}_p)$ is finite, there is an integer m such that

$$E(\mathbb{F}_p) \subset E(\mathbb{F}_p)[m] = \{x \in E(\mathbb{F}_p) : mx = 0\}.$$

It is a nontrivial fact[2] that for any elliptic curve over any field K, the m- 2
torsion subgroup $E(K)[m]$ is a subgroup of $\mathbb{Z}/m \times \mathbb{Z}/m$. For example, when $K \subset \mathbb{C}$ this follows from the fact that

$$\begin{aligned} E(K)[m] \subset E(\mathbb{C})[m] &= (\mathbb{R}/\mathbb{Z} \oplus \mathbb{R}/\mathbb{Z})[m] \\ &= \left(\frac{1}{m}\mathbb{Z}\right)/\mathbb{Z} \oplus \left(\frac{1}{m}\mathbb{Z}\right)/\mathbb{Z} = \mathbb{Z}/m \times \mathbb{Z}/m. \end{aligned}$$

To finish the proof, we show using elementary group theory that any subgroup of $\mathbb{Z}/m \times \mathbb{Z}/m$ can be generated by two elements, by e.g., counting ℓ-torsion for each prime ℓ. □

Theorem 11.1.4 (Hasse). *The cardinality of $E(\mathbb{F}_p)$ is bounded as follows:*

$$|\#E(\mathbb{F}_p) - (p + 1)| < 2\sqrt{p},$$

and every possibility for $\#E(\mathbb{F}_p)$ occurs.

[3] 3

Elliptic curves over finite fields are useful for much more than just computational applications. As we will see in Chapter **??**, a key step in the proof of Fermat's Last Theorem involves considering an elliptic curve $y^2 = x^3 + ax + b$ over $\mathbb{Q}$, and showing that a certain generating function whose coefficients encode $\#E(\mathbb{F}_p)$ (and other related information), for all but finitely many p, has good transformation properties.

11.2 Factorization

In 1987, Hendrik Lenstra published the landmark paper [Len87] that describes and analyzes the Elliptic Curve Method (ECM), which is a powerful algorithm for factoring integers using elliptic curves. Lenstra's method is also described in [ST92, §IV.4], [Dav99, §VIII.5], and [Coh93, §10.3].

[2] Give reference to Silverman?

[3] Give references.

Lenstra's algorithm is well-suited for finding "medium sized" factors of an integer N, which today means 10 to 20 decimal digits. The ECM method is not directly useful for factoring RSA challenge numbers (see Section 3.1.3), but surprisingly it is used in intermediate steps of some the algorithms that are used for hunting for such factorizations. Implementation of ECM typically requires little memory. Lenstra's discovery of ECM was inspired by Pollard's $(p-1)$-method, which we will describe in Section 11.2.1 below.

Lenstra

11.2.1 Pollard's $(p-1)$-Method

Definition 11.2.1 (Power-smooth). Let B be a positive integer. A positive integer n is *B-power smooth* if all prime powers dividing n are less than or equal to B.

Thus 30 is 7-power smooth and 5-power smooth, but 4 is not 2-power smooth.

Let N be a positive integer that we wish to factor. We use the Pollard $(p-1)$-method to look for a nontrivial factor of N as follows. First we choose a positive integer B, usually $\leq 10^6$ in practice. Suppose that there is a prime divisor p of N such that $p-1$ is B-power smooth. We try to find p computationally using the following strategy. If $a > 1$ is an integer not divisible by p then by Theorem 3.3.14,

$$a^{p-1} \equiv 1 \pmod{p}.$$

Letting $m = \operatorname{lcm}(1, 2, 3, \ldots, B)$, our assumption that $p-1$ is B-power smooth implies that $p-1 \mid m$, so

$$a^m \equiv 1 \pmod{p}.$$

Thus

$$p \mid \gcd(a^m - 1, N) > 1.$$

If $\gcd(a^m - 1, N) < N$ also then $\gcd(a^m - 1, N)$ is a nontrivial factor of N. If $\gcd(a^m - 1, N) = N$, then $a^m \equiv 1 \pmod{q^r}$ for every prime power divisor q^r of N. In this case, repeat the above steps but with a smaller choice of B or possibly a different choice of a. Also, check from the start whether or not N is not a perfect power M^r, and if so replace N by M.

For fixed B, this algorithm usually splits N when N is divisible by a prime p such that $p-1$ is B-power smooth. Only approximately 15% of primes p in the interval from 10^{15} and $10^{15} + 10000$ are such that $p-1$ is 10^6 power-smooth, so the Pollard method with $B = 10^6$ already fails nearly 85% of the time at finding 15-digit primes in this range. We will not analyze Pollard's method further, since it was mentioned here only to set the stage for the ECM.

The following examples illustrate the Pollard $(p-1)$-method.

Example 11.2.2. In this example, Pollard works perfectly. Let $N = 5917$. We try to use the Pollard $p-1$ method with $B = 5$ to split N. We have $m = \mathrm{lcm}(1, 2, 3, 4, 5) = 60$; taking $a = 2$ we have

$$2^{60} - 1 \equiv 3416 \pmod{5917}$$

and

$$\gcd(2^{60} - 1, 5917) = \gcd(3416, 5917) = 61,$$

so 61 is a factor of 5917.

Example 11.2.3. In this example, we replace B by larger integer. Let $N = 779167$. With $B = 5$ and $a = 2$ we have

$$2^{60} - 1 \equiv 710980 \pmod{779167},$$

and $\gcd(2^{60} - 1, 779167) = 1$. With $B = 15$, we have $m = \mathrm{lcm}(1, 2, \ldots, 15) = 360360$,

$$2^{360360} - 1 \equiv 584876 \pmod{779167},$$

and

$$\gcd(2^{360360} - 1, N) = 2003,$$

so 2003 is a nontrivial factor of 779167.

Example 11.2.4. In this example, we replace B by a smaller integer. Let $N = 4331$. Suppose $B = 7$, so $m = \mathrm{lcm}(1, 2, \ldots, 7) = 420$,

$$2^{420} - 1 \equiv 0 \pmod{4331},$$

and $\gcd(2^{420} - 1, 4331) = 4331$, so we do not obtain a factor of 4331. If we replace B by 5, Pollard's method works:

$$2^{60} - 1 \equiv 1464 \pmod{4331},$$

and $\gcd(2^{60} - 1, 4331) = 61$, so we split 4331.

Example 11.2.5. In this example, $a = 2$ does not work, but $a = 3$ does. Let $N = 187$. Suppose $B = 15$, so $m = \mathrm{lcm}(1, 2, \ldots, 15) = 360360$,

$$2^{360360} - 1 \equiv 0 \pmod{187},$$

and $\gcd(2^{360360} - 1, 187) = 187$, so we do not obtain a factor of 187. If we replace $a = 2$ by $a = 3$, then Pollard's method works:

$$3^{360360} - 1 \equiv 66 \pmod{187},$$

and $\gcd(3^{360360} - 1, 187) = 11$. Thus $187 = 11 \cdot 17$.

11.2.2 Motivation for the Elliptic Curve Method

Fix a positive integer B. If $N = pq$ with p and q prime and $p-1$ and $q-1$ are not B-power smooth, then the Pollard $(p-1)$-method is unlikely to work. For example, let $B = 20$ and suppose that $N = 59 \cdot 101 = 5959$. Note that neither $59 - 1 = 2 \cdot 29$ nor $107 - 1 = 2 \cdot 53$ is B-power smooth. With $m = \mathrm{lcm}(1, 2, 3, \ldots, 20) = 232792560$, we have

$$2^m - 1 \equiv 5944 \pmod{N},$$

and $\gcd(2^m - 1, N) = 1$, so we do not find a factor of N.

As remarked above, the problem is that $p-1$ is not 20-power smooth for either $p = 59$ or $p = 101$. However, notice that $p - 2 = 3 \cdot 19$ is 20-power smooth. Lenstra's ECM replaces $\mathbb{F}_p^\times$, which has order $p - 1$, by the group of points on an elliptic curve E over $\mathbb{F}_p$. By Theorem 11.1.4,

$$\#E(\mathbb{F}_p) = p + 1 \pm s$$

for some nonnegative integer $s < 2\sqrt{p}$ and any s can occur. For example, if E is the elliptic curve

$$y^2 = x^3 + x + 54$$

over $\mathbb{F}_{59}$ then by enumerating points one sees that $E(\mathbb{F}_{59})$ is cyclic of order 57 (every abelian group of order 57 is cyclic). The set of numbers $59+1\pm s$ for $s \leq 15$ contains 14 numbers that are B-power smooth for $B \leq 20$. For example, $60 = 59 + 1 + 0$ is 5-power smooth and $70 = 59 + 1 + 10$ is 7-power smooth.

11.2.3 The Elliptic Curve Method

The following description of the ECM algorithm is taken from [Len87], with slight changes to the notation and wording.

> The new method is obtained from Pollard's $(p-1)$-method by replacing the multiplicative group $\mathbb{F}_p^\times$ by the group of points on a random elliptic curve. To find a non-trivial divisor of an integer $N > 1$, one begins by selecting an elliptic curve E over $\mathbb{Z}/N$, a point P on E with coordinates in $\mathbb{Z}/N$, and an integer $m = \mathrm{lcm}(2, 3, \ldots, B)$. Using the addition law of the curve, one next calculates the multiple $m \cdot P$ of P. One now hopes that there is a prime divisor p of N for which $m \cdot P$ and the neutral element $\mathcal{O}$ of the curve become the same modulo p; if E is given by a Weierstrass equation $y^2 = x^3 + ax + b$, with $\mathcal{O} = (0 : 1 : 0)$, then this is equivalent to the third coordinate of $m \cdot P$ being divisible by p. Hence one hopes to find a non-trivial factor of N by calculating the greatest common divisor of this third coordinate with m.

If the above algorithm fails with a specific elliptic curve E, there is an option that is unavailable with Pollard's $(p-1)$-method. We may repeat the above algorithm with a different choice of E. The number of points on E over $\mathbb{Z}/p$ is of the form $p + 1 - t$ for some t with $|t| < 2\sqrt{p}$, and the algorithm is likely to succeed if $p + 1 - t$ is B-power-smoth.

Suppose that $P = (x_1, y_1)$ and $Q = (x_2, y_2)$ are nonzero points on an elliptic curve $y^2 = x^3 + ax + b$ and that $P \neq \pm Q$. Let $\lambda = (y_1 - y_2)/(x_1 - x_2)$ and $\nu = y_1 - \lambda x_1$. Recall from Section 10.2.8 the explicit formula for computing $P + P$ and also that $P + Q = (x_3, y_3)$ where

$$x_3 = \lambda^2 - x_1 - x_2 \qquad \text{and} \qquad y_3 = -\lambda x_3 - \nu.$$

We try to compute mP using the powering algorithm from Section 3.5.2. If at some step we can not compute $2^i P + 2^j P$ because we can not compute the inverse modulo N of $x_1 - x_2$, or we can not compute $2^i P$ because we can not compute the inverse of y_1 modulo N, then we compute the gcd of N and $x_1 - x_2$ or y_1. With luck, this gcd is a nontrivial divisor of N.

11.2.4 Examples

For simplicity, we use an elliptic curve of the form

$$y^2 = x^3 + ax + 1,$$

which has the point $P = (0, 1)$ already on it.

We factor $N = 5959$ using ECM, then we factor a much larger integer. Let

$$m = \mathrm{lcm}(1, 2, \ldots, 20) = 232792560 = 1101111000000010000111111 0000_2,$$

where x_2 means x is written in binary. First we choose $a = 1201$ at random and consider $y^2 = x^3 + 1201x + 1$ over $\mathbb{Z}/5959$. Using the formula for $P + P$ from Section 10.2.8 (implemented on a computer) we compute $2^i \cdot P = 2^i \cdot (0, 1)$ for $i \in B = \{4, 5, 6, 7, 8, 13, 21, 22, 23, 24, 26, 27\}$. Then $\sum_{i \in B} 2^i P = mP$. It turns out that during no step of this computation does a number not coprime to 5959 appear in any denominator, so we do not split N using $a = 1201$. Next we try $a = 389$ and at some stage in the computation we have to add $P = (2051, 5273)$ and $Q = (637, 1292)$. When computing the group law explicitly we try to compute $\lambda = (y_1 - y_2)/(x_1 - x_2)$ in $(\mathbb{Z}/5959)^\times$, but fail since $x_1 - x_2 = 1414$ and $\gcd(1414, 5959) = 101$. We thus find a nontrivial factor 5959 of 101.

11.2.5 A Conceptual Connection

Let N be a positive integer and for simplicity of exposition assume that $N = p_1 \cdots p_r$ with the p_i distinct primes. Recall from Section 3.4.1 that there is an isomorphism

$$f : (\mathbb{Z}/N)^\times \longrightarrow (\mathbb{Z}/p_1)^\times \times \cdots \times (\mathbb{Z}/p_r)^\times.$$

When using Pollard's method, we choose an $a \in (\mathbb{Z}/N)^\times$, compute a^m, then compute $\gcd(a^m - 1, N)$. This gcd is divisible exactly by the primes p_i such that $a^m \equiv 1 \pmod{p_i}$. To reinterpret Pollard's method using the above isomorphism, let $(a_1, \ldots, a_r) = f(a)$. Then $(a_1^m, \ldots, a_r^m) = f(a^m)$, and the p_i that divide $\gcd(a^m - 1, N)$ are exactly the p_i such that $a_i^m = 1$.

These are, in turn, the primes p_i such that $p_i - 1$ is B-power smooth, where $m = \mathrm{lcm}(1, \ldots, m)$.

From this point of view, the only significant difference between Pollard's method and ECM is that the isomorphism f is replaced by an isomorphism

$$g_a : E_a(\mathbb{Z}/N)^\times \to E_a(\mathbb{Z}/p_1) \times \cdots \times E_a(\mathbb{Z}/p_r)$$

where E_a is defined by $y^2 = x^3 + ax + 1$, and the a of Pollard's method is replaced by the point $P = (0 : 1 : 1)$. Here $E_a(\mathbb{Z}/N)^\times$ is the *group* of elements in

$$\mathbb{P}^2(\mathbb{Z}/N) = \frac{\{(x : y : z) : x, y, z \in \mathbb{Z}/N \text{ and } \gcd(x, y, z) = 1\}}{(\text{scalar multiplication by } (\mathbb{Z}/N)^\times)}$$

that satisfy $y^2 z = x^3 + axz^2 + z^3$. The map g_a is defined by reducing $(x : y : z)$ modulo p_i for each i. When carrying out the ECM we compute mP and if some of the component of $g_a(mP)$ are zero, but others are nonzero, we find a nontrivial factor of N by taking the gcd of N and the third component of mP. The advantage of ECM is that for a fixed m we can carry out this process for many different choices of a, each time increasing the chances that we will split off "medium sized" factors of N.

11.3 Cryptography

In this section we discuss analogues of Diffie-Hellman and RSA for elliptic curves. We then discuss how the elliptic curve cryptosystem used in version 2 of the Microsoft Digital Rights Management (MS-DRM) system was cracked.

11.3.1 Elliptic Curve Analogues of RSA and Diffie-Hellman

The Diffie-Hellman key exchange from Section 4.1 works well on an elliptic curve with no serious modification. Michael and Nikita agree on a secret key as follows:

1. Michael and Nikita agree on a prime p, an elliptic curve E over $\mathbb{Z}/p$, and a point $P \in E(\mathbb{Z}/p)$.

2. Michael secretely chooses a random m and sends mP.

3. Nikita secretely chooses a random n and sends nP.

4. The secrete key is nmP, which both Michael and Nikita can compute.

Presumably, an adversary can not compute nmP without solving the discrete logarithm problem (see Problem 4.1.2 and Section 11.3.3 below) in $E(\mathbb{Z}/p)$. For well-chosen E, P, and p experience suggests that the discrete logarithm problem in $E(\mathbb{Z}/p)$ is much more difficult than the discrete logarithm problem in $(\mathbb{Z}/p)^{\times}$.

There is an analogue for elliptic curves of the RSA cryptosystem of Section 4.2, but the author has never heard of anyone actually using it. Nikita sets up an RSA-elliptic curve public key, as follows:

1. Nikita secretely chooses primes p and q, and lets $N = pq$.

2. Nikita chooses an elliptic curve E over $\mathbb{Z}/N$ and considers the group $E(\mathbb{Z}/N)$ (see Section 11.2.5 for the meaning of $E(\mathbb{Z}/N)$).

3. Since Nikita knows p and q, she can use a sophisticated polynomial time algorithm of Schoof, Elkies, and Atkin (see [4]) to compute 4

$$m = \#E(\mathbb{Z}/N) = \#E(\mathbb{Z}/p) \cdot \#E(\mathbb{Z}/q).$$

4. Nikita chooses a random integer e between 1 and $m-1$ that is coprime to m. She lets d be the inverse of e modulo m.

5. To encrypt a message to Nikita, Michael encodes the message as a point $P \in E(\mathbb{Z}/N)$, then sends eP. To decrypt, Nikita computes $d(eP) = (de)P = P$.

This is at best no more secure than RSA, since factoring N breaks the cryptosystem, which probably explains why it is so unpopular.

[4] Give reference.

11.3.2 The ElGamal Cryptosystem and Microsoft Digital Rights Management

This section is about the ElGamal cryptosystem, which works well on an elliptic curves. It is used in version 2 of the Microsoft Digital Rights Management (MS-DRM) system.

This section draws on a paper by a hacker named Beale Screamer who cracked version 2 of MS-DRM.

The elliptic curve used in MS-DRM is an elliptic curve over the finite field $k = \mathbb{F}_p$, where

$$p = 785963102379428822376694789446897396207498568951.$$

As Beale Screamer remarks, this modulus has high nerd appeal because in hexadecimal it is

$$89\text{ABCDEF}012345672718281831415926141424\text{F}7,$$

which includes counting in hexadecimal, and digits of e, π, and $\sqrt{2}$. The Microsoft elliptic curve E is

$$\begin{aligned} y^2 = x^3 &+ 317689081251325503476317476413827693272746955927x \\ &+ 79052896607878758718120572025718535432100651934. \end{aligned}$$

We have

$$\#E(k) = 785963102379428822376693024881714957612686157429,$$

and the group $E(k)$ is cyclic with generator

$$\begin{aligned} B = (&771507216262649826170648268565579889907769254176, \\ &390157510246556628525279459266514995562533196655). \end{aligned}$$

Our heros Nikita and Michael love to share digital music when they are not out thwarting terrorists. When Nikita installed Microsoft's content rights management software on her laptop, it generated a private key

$$n = 670805031139910513517527207693060456300217054473,$$

which it hid in bits and pieces of files (e.g., `blackbox.dll`, `v2ks.bla`, and `IndivBox.key`). In order for Nikita to play Juno Reactor's latest hit `juno.wma`, her web browser contacts a Microsoft rights management partner. After Nikita sends her credit card number, the rights management partner sends her a license to play `juno.wma`.

As we will see below, the license file was created using the ElGamal public-key cryptosystem in the group $E(k)$. Nikita can now use her license file to unlock `juno.wma`. However, when she shares both `juno.wma` and the license file with Michael, he is frustrated because even with the license his laptop still does not play `juno.wma`. This is because Michael's laptop does not know Nikita's laptop's private key (the integer n above), so Michael's laptop can not decrypt the license file.

`juno.wma`

11.3.3 The Elliptic Curve Discrete Logarithm Problem

Definition 11.3.1. If E is an elliptic curve over $\mathbb{F}_p$ and B is a point on E, then the *discrete log problem* on E to the base B is the following problem: given a point $P \in E$ such that $P = mB$ for some m, find an integer n such that $P = nB$.

For example, let E be the elliptic curve given by $y^2 = x^3 + x + 1$ over the field $\mathbb{F}_7$. We have

$$E(\mathbb{F}_7) = \{\mathcal{O}, (2,2), (0,1), (0,6), (2,5)\}.$$

If $B = (2,2)$ and $P = (0,6)$, then $3B = P$, so $n = 3$ is a solution to the discrete logarithm problem.

When p is large, the discrete logarithm problem on an elliptic curve E over $\mathbb{F}_p$ is conjectured to be "very difficult", except in two special cases: if $\#E(\mathbb{F}_p)$ is "smooth" (i.e., a product of small primes) or E is "supersingular" (i.e., $\#E(\mathbb{F}_p) = p + 1$).[5] The Microsoft curve has neither of these 5
deficiencies, so we expect that the discrete logarithm on that curve is difficult. Beale Screamer does not solve the discrete logarithm on E; this is not how he circumvents MS-DRM.

11.3.4 ElGamal

The ElGamal public-key cryptosystem lends itself well to implementation in the group $E(\mathbb{F}_p)$. To illustrate ElGamal, we describe how Nikita would set up an ElGamal cryptosystem that anyone could use to encrypt messages for her. Nikita chooses a prime p, an elliptic curve E over $\mathbb{F}_p$, and a point $B \in E(\mathbb{F}_p)$, and publishes p, E, and B. She also chooses a random integer n,

[5] Find references.

which she keeps secret, and publishes nB. Her public key is the four-tuple (p, E, B, nB).

Suppose Michael wishes to encrypt a message for Nikita. If the message is encoded as an element $P \in E(\mathbb{F}_p)$, Michael computes a random integer r and the points rB and $P + r(nB)$ on $E(\mathbb{F}_p)$. Then P is encrypted as the pair $(rB, P + r(nB))$. To decrypt the encrypted message, Nikita multiplies rB by her secret key n to find $n(rB) = r(nB)$, then subtracts this from $P + r(nB)$ to obtain

$$P = P + r(nB) - r(nB).$$

Example 11.3.2. Nikita's license files contains the pair of points $(rB, P + r(nB))$, where

$$\begin{aligned} rB = (&179671003218315746385026655733086044982194424660,\\ &697834385359686368249301282675141830935176314718) \end{aligned}$$

and

$$\begin{aligned} P + r(nB) = (&137851038548264467372645158093004000343639118915,\\ &1108485892286762240572292302235808150242224875699). \end{aligned}$$

Nikita's laptop loads the secret key

$$n = 670805031139910513517527207693060456300217054473$$

into memory and computes

$$\begin{aligned} n(rB) = r(nB) = (&328901393518732637577115650601768681044040715701,\\ &586947838087815993601350565488788846203887988162). \end{aligned}$$

It then subtracts this from $P + r(nB)$ to obtain

$$\begin{aligned} P = (&14489646124220757767,\\ &669337780373284096274895136618194604469696830074). \end{aligned}$$

The x coordinate 14489646124220757767 is the content key that unlocks `juno.wma`.

If Nikita knew the private key n that her laptop generated, she could compute P herself and unlock `juno.wma` and share her music with Michael. Beale Screamer found a weakness in Microsoft's system that let him find n:

> "These secret keys are stored in linked lists ... interspersed with the code in the library. The idea is that they can be read by that library, used internally by that library, and never communicated outside the library. Since the `IndivBox.key` file is shuffled in a random way for each client, these keys would be extremely difficult to extract from the file itself. Fortunately, we don't have to: these keys are part of the object state that is maintained by this library, and since the offset within this object of these secret keys is known, we can let the library itself extract the

> secret keys! The code for this simply loads up the 'black box' library, has it initialize an instance of the object, and then reads the keys right out of that object. This is clearly a weakness in the code which can be corrected by the DRM software fairly easily, but for now it is the basis of our exploit."

Open Problem 11.3.3. *How can Microsoft store data on Nikita's laptop in such a way that Nikita can not access it, but Nikita's laptop can?*

11.3.5 Why Use Elliptic Curves?

There are several advantages to using elliptic curves in cryptography.

Elliptic curve based cryptosystems with relatively small key sizes are, under certain assumptions, provably as secure as "classical" cryptosystems like RSA with much larger key sizes. And size does matter. According to Dan Boneh of Stanford University, Microsoft may soon use an elliptic curve based cryptosystem during the installation of some of their products. It is unreasonable to ask a user to type in a license key with hundreds of digits; using an elliptic curve system, they can ask the user to type in a much smaller key instead.

EXERCISES

11.1 Let $N = pq$ be a product of distinct odd primes and let $a, b \in \mathbb{Z}/N$ be such that $4ba^3 + 27b^2 \neq 0$. Let E be the elliptic curve defined by $y^2 = x^3 + ax + b$. Prove that reduction modulo p and modulo q induces an isomorphism $E(\mathbb{Z}/N) \to E(\mathbb{Z}/p) \times E(\mathbb{Z}/q)$. (See Section 11.2.5 for a discussion of the meaning of $E(\mathbb{Z}/N)$.)

12
Modular Forms and Elliptic Curves

Let E be an elliptic curve over $\mathbb{Q}$, so E is defined by an equation $y^2 = x^3 + ax + b$ with $a, b \in \mathbb{Q}$. It was recently proved by Andrew Wiles and others that all such elliptic curves are "modular", a result which provides a huge number of tools for studying elliptic curves over $\mathbb{Q}$. Two important consequences are that Fermat's Last Theorem is true, and that the conjecture of Birch and Swinnerton-Dyer about the rank of $E(\mathbb{Q})$ (see Chapter 13) involves objects that are defined.

In Section 12.1 we define modular forms, and in Section 12.2 we give a definition of what it means for an elliptic curve to be modular. Section 12.3 contains a brief discussion of how modularity of elliptic curves implies the truth of Fermat's Last Theorem.

There is a vast amount of literature about modular forms. See [Ser73, Ch. 7] for a first introduction, then look at the modern survey paper [DI95] for an excellent overview of the most important basic facts about modular forms along with a fairly complete bibliography.

12.1 Modular Forms

The complex *upper half plane* is the set

$$\mathfrak{h} = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}.$$

A *holomorphic function* $f : \mathfrak{h} \to \mathbb{C}$ is a function such that for all $z \in \mathfrak{h}$ the derivative

$$f'(z) = \lim_{h \to 0} \frac{f(z+h) - f(z)}{h}$$

exists. Let $\mathcal{H}$ denote the complex vector space of holomorphic functions on $\mathfrak{h}$.

Holomorphicity is a very strong condition because $h \in \mathbb{C}$ can approach 0 in many ways. For example, if $f(z)$ is holomorphic, then all derivatives $f^{(n)}(z)$ automatically exist, and $f(z)$ converges to its Taylor expansion in a neighborhood of any point.

Recall that $\mathrm{SL}_2(\mathbb{Z})$ denotes the group of 2×2 integer matrices with determinant 1. The linear fractional transformation induced by $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$ is

$$\gamma(z) = \frac{az+b}{cz+d}.$$

The group $\mathrm{SL}_2(\mathbb{Z})$ acts on the right on $\mathcal{H}$ by pre-composition:

$$f(z) \mapsto f(\gamma(z)).$$

The space of *holomorphic differentials on* $\mathfrak{h}$ is the complex vector space of expressions

$$\Omega = \{f(z)dz : f \text{ is a holomorphic function on } \mathfrak{h}\}.$$

There is a bijection between the holomorphic functions on $\mathfrak{h}$ and the holomorphic differentials on $\mathfrak{h}$ given by $f(z) \mapsto f(z)dz$ (the inverse is $\omega \mapsto \omega/dz$). The group $\mathrm{SL}_2(\mathbb{Z})$ acts on Ω in a different and more interesting way than it acts on $\mathcal{H}$. For $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ and $f(z)dz \in \Omega$, let

$$(f(z)dz)|_\gamma = f(\gamma(z))d(\gamma(z)).$$

Remark 12.1.1. The quotient rule from calculus and that $\det(\gamma) = 1$ imply that $d(\gamma(z)) = (cz+d)^{-2}dz$. Thus under the bijection between Ω and $\mathcal{H}$, the action of $\mathrm{SL}_2(\mathbb{Z})$ on Ω corresponds to the action of $\mathrm{SL}_2(\mathbb{Z})$ on $\mathcal{H}$ given by

$$f(z)|_\gamma = f(\gamma(z))(cz+d)^{-2}.$$

For any positive integer N, consider the subgroup

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \,:\, N \mid c \right\} \subset \mathrm{SL}_2(\mathbb{Z}).$$

Let $\Omega(\Gamma_0(N))$ be the subspace of Ω of functions f such that $f(z)dz$ is fixed by every element of $\Gamma_0(N)$. If $f(z)dz \in \Omega(\Gamma_0(N))$, then since $\left(\begin{smallmatrix} 1 & h \\ 0 & 1 \end{smallmatrix}\right) \in \Gamma_0(N)$ for some integer $h > 0$ (in fact, $h = 1$), we have $f(z+h)dz = f(z)dz$, so $f(z+h) = f(z)$.

Proposition 12.1.2. *The holomorphic function $q_h(z) = e^{2\pi i z/h}$ maps the vertical strip*

$$V = \{z \in \mathfrak{h} : 0 \leq \mathrm{Re}(z) < h\}$$

bijectively onto the punctured open unit disk $D = \{z \in \mathbb{C} : 0 < |z| < 1\}$. If $f : \mathfrak{h} \to \mathbb{C}$ is a function that satisfies $f(z+h) = f(z)$, then there is a unique function $F : D \to \mathbb{C}$ such that $f(z) = F(q_h(z))$.

Proof. We will not prove that e is holomorphic, because this is a standard part of any complex analysis course. If $z = x + iy \in V$, then

$$e^{2\pi i z/h} = e^{2\pi i(x+iy)/h} = e^{-2\pi y/h} e^{2\pi i x/h}$$

is in D since $y > 0$. Every element of D is uniquely of the form $e^{-2\pi y/h}e^{2\pi ix/h}$ for $y/h > 0$ and $0 \leq x/h < 1$, so q_h is a bijection.

For $w \in D$ let $F(w) = f(q_h^{-1}(w))$, where $q_h^{-1} : D \to V$ is the inverse of q_h. Then for $z \in V$, we have $F(q_h(z)) = f(q_h^{-1}(q_h(z))) = f(z)$ as required. Since $f(z) = f(z+1)$, we have $F(q_h(z)) = f(z)$ for all $z \in \mathfrak{h}$. □

Suppose $f \in \mathcal{H}$ satisfies $f(z+h) = f(z)$ for some positive integer h. Then $f(z)$ is *holomorphic at infinity* if the function $F(q_h)$ of Proposition 12.1.2 on $D \subset \mathbb{C}$ extends to a holomorphic function at 0. If this extension (which is necessarily unique) is 0 at 0, we say that f *vanishes at infinity.* If f is holomorphic at infinity, then $F(q_h)$ has a Taylor expansion (it is a theorem in complex analysis that holomorphic functions converge to their Taylor expansions) and there is a neighborhood of infinity such that

$$f = \sum_{n=0}^{\infty} a_n q_h^n$$

for complex numbers a_n. This expansion is called the *q-expansion of $f(z)$ at infinity.*

Definition 12.1.3 (Modular Forms). The vector space of *modular forms (of weight 2) for* $\Gamma_0(N)$ is the subspace $M_2(\Gamma_0(N))$ of $\mathcal{H}$ of holomorphic function $f : \mathfrak{h} \to \mathbb{C}$ such that

1. $f(z)dz \in \Omega(\Gamma_0(N))$
2. For every $\alpha \in \mathrm{SL}_2(\mathbb{Z})$, the *function* $(f(z)dz)|_\alpha/dz$ is holomorphic at infinity.

It takes some work to see that the second condition in the definition makes sense.

Lemma 12.1.4. *If $\alpha \in \mathrm{SL}_2(\mathbb{Z})$ and $\omega \in \Omega(\Gamma_0(N))$ then $\omega|_\alpha$ is fixed by $\alpha^{-1}\gamma\alpha$ for any $\gamma \in \Gamma_0(N)$.*

Proof. We have
$$(\omega|_\alpha)|_{\alpha^{-1}\gamma\alpha} = \omega|_{\gamma\alpha} = \omega|_\alpha.$$
□

If $\alpha = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ then there exists h such that if $t_h = \left(\begin{smallmatrix} 1 & h \\ 0 & 1 \end{smallmatrix}\right)$, then

$$\gamma = \alpha t_h \alpha^{-1} = \begin{pmatrix} -ach+1 & ha^2 \\ -hc^2 & hac+1 \end{pmatrix} \in \Gamma_0(N).$$

Thus $t_h = \alpha^{-1}\gamma\alpha$ with $\gamma \in \Gamma_0(N)$, so by Lemma 12.1.4, $(\omega|_\alpha)|_{t_h} = \omega|_\alpha$ so $g(z) = \omega|_\alpha/dz$ satisfies $g(z+h) = g(z)$. Thus g has a Fourier expansion at ∞ and condition 2 makes sense.

Remark 12.1.5. It is unfortunate that modular forms are actually functions instead of differential forms, but this is standard in the literature.

Definition 12.1.6. The subspace $S_2(\Gamma_0(N))$ of *cusp forms* is the subspace of elements $f \in M_2(\Gamma_0(N))$ such that the function $(f(z)dz)|_\alpha/dz$ vanishes at infinity for all $\alpha \in \mathrm{SL}_2(\mathbb{Z})$.

The cusp forms correspond to the differentials that are holomorphic even at the points at infinity, in the following sense. Letting $q = e^{2\pi i z}$, we have $\frac{dq}{q} = dz$, so if $f(q) = \sum_{n=0}^{\infty} a_n q^n$, then the differential

$$f(z)dz = f(q)\frac{dq}{q} = \left(\frac{a_0}{q} + a_1 + a_2 q + a_3 q^2 + \cdots\right) dq$$

is "holomorphic at infinity" if and only if $a_0 = 0$.

Remark 12.1.7. The condition that $(f(z)dz)|_\alpha/dz$ have a nice property at infinity for all $\alpha \in \mathrm{SL}_2(\mathbb{Z})$ probably seems ad hoc. It is motivated by the following geometric observation. The quotient of $\mathfrak{h}$ by the action of $\Gamma_0(N)$ is a noncompact Riemann surface $Y_0(N)$ (it is missing a finite set set of points). Elements of $\Omega(\Gamma_0(N))$ correspond to differentials on $Y_0(N)$. Differentials on noncompact Riemann surfaces are not well behaved; for example, the space of holomorphic differentials will not be finite dimensional. The differentials on $Y_0(N)$ that extend to holomorphic differentials on the compactification $X_0(N)$ are exactly the elements of $S_2(\Gamma_0(N))$. This is a finite dimensional space with dimension equal to the genus (number of holes) of the Riemann surface $X_0(N)$.

12.1.1 Examples

The following theorem is proved using techniques from algebraic geometry:

Theorem 12.1.8. *The complex vector space $S_2(\Gamma_0(N))$ has finite dimension:*

$$\dim S_2(\Gamma_0(N)) = 1 + \frac{\mu}{12} - \frac{\nu_2}{4} - \frac{\nu_3}{3} - \frac{\nu_\infty}{2},$$

where

$$\mu = N \prod_{p\mid N} (1 + 1/p)$$

$$\nu_2 = \begin{cases} 0 & \text{if } 4 \mid N \\ \prod_{p\mid N}\left(1 + \left(\frac{-4}{p}\right)\right) & \text{otherwise} \end{cases}$$

$$\nu_3 = \begin{cases} 0 & \text{if } 2 \mid N \text{ or } 9 \mid N \\ \prod_{p\mid N}\left(1 + \left(\frac{-3}{p}\right)\right) & \text{otherwise} \end{cases}$$

$$\nu_\infty = \sum_{d\mid N} \varphi(\gcd(d, N/d)).$$

For example,

$$\dim_{\mathbb{C}} S_2(\Gamma_0(2)) = 1 + \frac{3}{12} - \frac{1}{4} - \frac{0}{3} - \frac{2}{2} = 0,$$

and

$$\dim_{\mathbb{C}} S_2(\Gamma_0(11)) = 1 + \frac{12}{12} - \frac{0}{4} - \frac{0}{3} - \frac{2}{2} = 1.$$

For the rest of this section, let $q(z) = e^{2\pi i z}$. The following basis were computed using algorithms implemented in [BCP97] that are beyond the scope of this book.

Example 12.1.9. The vector space $M_2(\Gamma_0(11))$ has basis

$$f_1 = 5 + 12q + 36q^2 + 48q^3 + 84q^4 + 72q^5 + 144q^6 + 96q^7 + \cdots$$
$$f_2 = q - 2q^2 - q^3 + 2q^4 + q^5 + 2q^6 - 2q^7 + \cdots .$$

and the subspace $S_2(\Gamma_0(11))$ of cusp forms has basis f_2.

The smallest N such that $S_2(\Gamma_0(N))$ has dimension bigger than 1 is $N = 22$. A basis for this space is

$$f_1 = q - q^3 - 2q^4 + q^5 - 2q^7 + \cdots ,$$
$$f_2 = q^2 - 2q^4 - q^6 + \cdots ,$$

Example 12.1.10. The space $S_2(\Gamma_0(43))$ has dimension 3 and basis

$$f_1 = q + 2q^5 - 2q^6 - 2q^7 + \cdots ,$$
$$f_2 = q^2 + q^3 - q^4 + 3q^5 - 3q^6 - q^7 + \cdots ,$$
$$f_3 = 2q^3 - q^4 + 4q^5 - 3q^6 - 2q^7 + \cdots$$

12.2 Modular Elliptic Curves

Let E be an elliptic curve defined by a Weierstrass equation $y^2 = x^3 + ax + b$ with $a, b \in \mathbb{Z}$. (If $a, b \in \mathbb{Q}$, then the equation can be transformed into one with $a, b \in \mathbb{Z}$; see Exercise 2.) For each prime $p \nmid \Delta = -16(4a^3 + 27b^2)$, set

$$a_p = p + 1 - \#E(\mathbb{Z}/p\mathbb{Z}).$$

Definition 12.2.1 (Modular). Let $N = |\Delta|$ be the absolute value of the discriminant of $y^2 = x^3 + ax + b$ (with $a, b \in \mathbb{Z}$). Then the elliptic curve E defined by $y^2 = x^3 + ax + b$ is *modular* if there exists a cuspidal modular form

$$f(z) = \sum_{n=1}^{\infty} b_n q^n \in S_2(\Gamma_0(N))$$

such that $b_p = a_p$ for all $p \nmid \Delta$.

At first glance, modularity appears to be a bizarre and unlikely property for an elliptic curve to have. Yutaka Taniyama and Goro Shimura first suggested in 1955 that every elliptic curve is modular, but mathematicians were initially dubious. Andre Weil later gave significant theoretical evidence for the conjecture. Motivated by a deep connection between this conjecture and Fermat's last theorem, Andrew Wiles proved enough of the conjecture to deduce Fermat's last theorem. A full proof of the conjecture was finally completed in 1999, and it is one of the crowning achievements of number theory.

Theorem 12.2.2 (Breuil, Conrad, Diamond, Taylor, Wiles). *Every elliptic curve over $\mathbb{Q}$ is modular.*

Wiles

12.3 Fermat's Last Theorem

A huge amount of number theory has been motivated by attempts by number theorists to prove "Fermat's Last Theorem". This is a conjecture that was made in the 1600s and was finally proved over 300 years later.

Theorem 12.3.1 (Wiles [Wil95]). *Let $n > 2$ be an integer. If $a, b, c \in \mathbb{Z}$ and*

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

then $abc = 0$.

The proof generated an immense amount of excitement, which is illustrated in this famous mailing list post.

```
From K.C.Rubin@newton.cam.ac.uk Wed Jun 23 02:53:28 1993
Date: Wed, 23 Jun 93 10:50 BST
From: K.C.Rubin@newton.cam.ac.uk
Subject: big news

Andrew Wiles just announced, at the end of his 3rd lecture here,
that he has proved Fermat's Last Theorem.  He did this by proving
that every semistable elliptic curve over Q (i.e. square-free
conductor) is modular.  The curves that Frey writes down, arising
from counterexamples to Fermat, are semistable and by work of
Ribet they cannot be modular, so this does it.

It's an amazing piece of work.

Karl

From K.A.Ribet@newton.cam.ac.uk Wed Jun 23 05:40:01 1993
Date: Wed, 23 Jun 93 13:36 BST
From: K.A.Ribet@newton.cam.ac.uk
To: nts_local@math.berkeley.edu
Subject: announcement of Taniyama conjecture

I imagine that many of you have heard rumours about Wiles's
announcement a few hours ago that he can prove Taniyama's conjecture
for semistable elliptic curves over Q.  This case of the Taniyama
conjecture implies Fermat's Last Theorem, in view of the result
that I proved a few years ago.  (I proved that the "Frey elliptic
curve" constructed from a possible solution to Fermat's equation
cannot be modular, i.e., satisfy Taniyama's Conjecture.  On the
other hand, it is easy to see that it is semistable.)

Here is a brief summary of what Wiles said in his three lectures.

The method of Wiles borrows results and techniques from lots and lots
of people.  To mention a few: Mazur, Hida, Flach, Kolyvagin, yours
truly, Wiles himself (older papers by Wiles), Rubin...  The way he does
```

```
it is roughly as follows.  Start with a mod p representation of the
Galois group of Q which is known to be modular.  You want to prove that
all its lifts with a certain property are modular.  This means that the
canonical map from Mazur's universal deformation ring to its "maximal
Hecke algebra" quotient is an isomorphism.  To prove a map like this is
an isomorphism, you can give some sufficient conditions based on
commutative algebra.  Most notably, you have to bound the order of a
cohomology group which looks like a Selmer group for Sym^2 of the
representation attached to a modular form.  The techniques for doing
this come from Flach; you also have to use Euler systems a la
Kolyvagin, except in some new geometric guise.

If you take an elliptic curve over Q, you can look at the
representation of Gal on the 3-division points of the curve.  If you're
lucky, this will be known to be modular, because of results of Jerry
Tunnell (on base change).  Thus, if you're lucky, the problem I
described above can be solved (there are most definitely some
hypotheses to check), and then the curve is modular.  Basically, being
lucky means that the image of the representation of Galois on
3-division points is GL(2,Z/3Z).

Suppose that you are unlucky, i.e., that your curve E has a rational
subgroup of order 3.  Basically by inspection, you can prove that if it
has a rational subgroup of order 5 as well, then it can't be
semistable.  (You look at the four non-cuspidal rational points of
X_0(15).)  So you can assume that E[5] is "nice." Then the idea is to
find an E' with the same 5-division structure, for which E'[3] is
modular.  (Then E' is modular, so E'[5] = E[5] is modular.)  You
consider the modular curve X which parametrizes elliptic curves whose
5-division points look like E[5].  This is a "twist" of X(5).  It's
therefore of genus 0, and it has a rational point (namely, E), so it's
a projective line.  Over that you look at the irreducible covering
which corresponds to some desired 3-division structure.  You use
Hilbert irreducibility and the Cebotarev density theorem (in some way
that hasn't yet sunk in) to produce a non-cuspidal rational point of X
over which the covering remains irreducible.  You take E' to be the
curve corresponding to this chosen rational point of X.

-ken ribet
```

It turned out that Wiles's original proof contained a substantial gap. Fortunately, he and Richard Taylor worked hard over many months and found a new argument that bridged the gap. Their heroic struggle is portrayed in the Nova documentary *The Proof* (see [Sin97] for a transcript).

We now sketch a link between Fermat's Last Theorem and modularity of elliptic curves. It is easy to reduce to the case when $n = \ell$ is a prime greater than 3 (see Exercise 5 to reduce the the case n prime). Suppose that

$$a^\ell + b^\ell = c^\ell$$

with $a, b, c \in \mathbb{Z}$ and $abc \neq 0$. By dividing out by any common factor, we may assume that $\gcd(a, b, c) = 1$. Then permuting (a, b, c), we may suppose that b is even and that $a \equiv 3 \pmod 4$.

Following Gerhard Frey and Yves Hellegouarch, consider the elliptic curve E over $\mathbb{Q}$ defined by

$$y^2 = x(x - a^\ell)(x + b^\ell).$$

This equation is not of the the usual form $y^2 = x^3 + \alpha x + \beta$, but by replacing x by $x - (-a^\ell + b^\ell)$ it is transformed into the form $y^2 = x^3 + \alpha x + \beta$.

Lemma 12.3.2. *The discriminant of E is $2^4(abc)^{2\ell}$.*

Proof. Elementary algebra shows that the discriminant Δ of E is

$$(a^{2\ell} b^{2\ell} 2^4) \cdot (a^\ell + b^\ell)^2.$$

(As a check, note that this expression is 0 if and only if $x(x - a^\ell)(x + b^\ell)$ has a multiple root.) Thus

$$\Delta = (a^{2\ell} b^{2\ell} 2^4) \cdot c^{2\ell} = 2^4 \cdot (abc)^{2\ell}.$$

as claimed. $\square$

Remark 12.3.3. If we take random a^ℓ and b^ℓ such that $a^\ell + b^\ell$ is not an ℓth power, then the discriminant of the corresponding curve is far from being of the special form $2^4(abc)^{2\ell}$. For example, suppose $a^\ell = 3^5$ and $b^\ell = 7^5$. Then $a^\ell + b^\ell = 2 \cdot 5^2 \cdot 11 \cdot 31$, and the discriminant of $y^2 = x(x - 3^5)(x + 7^5)$ is $2^6 \cdot 3^{10} \cdot 5^4 \cdot 7^{10} \cdot 11^2 \cdot 31^2$.

Suppose again that E is defined by $y^2 = x(x - a^\ell)(x + b^\ell)$ with (a, b, c) a counterexample to Fermat's conjecture, as above. As in Section 12.2, for each prime $p \nmid abc$, let

$$a_p = p + 1 - \#E(\mathbb{F}_p).$$

By a deep special case of Theorem 12.2.2 that was proved by Wiles and Richard Taylor, there is a cusp form

$$g = \sum_{n=1}^{\infty} b_n q^n \in S_2(\Gamma_0(N)),$$

where $N = |2^4(abc)^{2\ell}|$, such that $a_p = b_p$ for all primes $p \nmid 2abc$.

Ken Ribet [Rib90] used that the discriminant of E is a perfect ℓth power (away from 2) to deduce that $g \bmod \ell$ comes from a much lower level, in the following sense: there is a nonzero cusp form

$$h = \sum_{n=1}^{\infty} c_n q^n \in S_2(\Gamma_0(2))$$

such that

$$b_p \equiv c_p \pmod{\ell} \qquad \text{for all } p \nmid 2abc.$$

Theorem 12.1.8 implies that $\dim S_2(\Gamma_0(2)) = 0$, which is a contradiction since g is nonzero. Thus the elliptic curve $y^2 = x(x - a^\ell)(x + b^\ell)$ can not exist, and our assumption that a, b, c are a solution to $a^\ell + b^\ell = c^\ell$ is false.

EXERCISES

12.1 Let $S_2(\Gamma_0(N))$ denote the set of cuspidal modular forms of level N. Prove that $S_2(\Gamma_0(N))$ forms a $\mathbb{C}$-vector space under addition.

12.2 Suppose $y^2 = x^3 + ax + b$ with $a, b \in \mathbb{Q}$ defines an elliptic curve. Show that there is another equation $Y^2 = X^3 + AX + B$ with $A, B \in \mathbb{Z}$ whose solutions are in bijection with the solutions to $y^2 = x^3 + ax + b$. (Hint: Multiply both sides of $y^2 = x^3 + ax + b$ by a power of a common denominator, and "absorb" powers into x and y.)

12.3 (a) Use Theorems 12.1.8 and 12.2.2 to deduce that there is no elliptic curve $y^2 = x^3 + ax + b$ (with $a, b \in \mathbb{Z}$) that has discriminant ± 16.

(b) The point $(12, 36)$ lies on the elliptic curve $y^2 = x^3 - 432$. Use this fact and elementary algebra to find a rational solution (a, b) to $4a^3 + 27b^2 = -1$, and hence exhibit an elliptic curve over $\mathbb{Q}$ with discriminant 16.

12.4 One can prove that the function

$$f = q \prod_{n=1}^{\infty} (1 - q^n)^2 (1 - q^{11n})^2 = \sum_{n=1}^{\infty} a_n q^n$$

spans $S_2(\Gamma_0(11))$, and that the following three matrices generate the subgroup $\Gamma_0(11)$ of $\mathrm{SL}_2(\mathbb{Z})$:

$$S = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \qquad T = \begin{pmatrix} 3 & -2 \\ 11 & -7 \end{pmatrix} \qquad U = \begin{pmatrix} 4 & -3 \\ 11 & -8 \end{pmatrix}.$$

Using the above product expression for f, compute f to some large precision then give numerical evidence that $f(z)$ satisfies the defining equation for an element of $S_2(\Gamma_0(11))$.

12.5 Show that if Fermat's last theorem is true for prime exponents, then it is true for all exponents.

12.6 Let R be a ring. Say that Fermat's last theorem is false in R if there exists $x, y, z \in R$ and $n \in \mathbb{Z}$ with $n \geq 3$ such that $x^n + y^n = z^n$ and $xyz \neq 0$. For which prime numbers p is Fermat's last theorem false in the ring $\mathbb{Z}/p$?

13

The Birch and Swinnerton-Dyer Conjecture

This chapter is about a conjecture that Birch and Swinnerton-Dyer made in the 1960s on the ranks of elliptic curves.

First we discuss the congruent number problem, which is an ancient problem that goes back over one thousand years, and see how it is connected with the Birch and Swinnerton-Dyer conjecture.

13.1 The Congruent Number Problem

Definition 13.1.1 (Congruent Number). A nonzero rational number n is called a *congruent number* if $\pm n$ is the area of a right triangle with rational side lengths. Equivalently, n is a *congruent number* if the system of two equations

$$n = \frac{ab}{2} \qquad \text{and} \qquad a^2 + b^2 = c^2$$

has a solution with $a, b, c \in \mathbb{Q}$.

For example, 6 is the area of the right triangle with side lengths 3, 4, and 5, so 6 is a congruent number. Less obvious is that 5 is also a congruent number; it is the area of the right triangle with side lengths $3/2$, $20/3$, and $41/6$. It is nontrivial to prove that 1, 2, 3, and 4 are not congruent numbers. Here is a list of the congruent numbers up to 50:

$$5, 6, 7, 13, 14, 15, 20, 21, 22, 23, 24, 28, 29, 30, 31, 34, 37, 38, 39, 41, 45, 46, 47, \ldots$$

Every congruence class modulo 8 except 3 is represented in this list, which suggests that if $n \equiv 3 \pmod 8$ then n is not a congruent number. This is true for $n \leq 218$, but $n = 219$ is a congruent number congruent to 3 mod 8. Something very subtle is going on.

This is another example which hints at the subtlety of congruent numbers. The number 157 is a congruent number, and Zagier showed that the *simplest* rational right triangle with area 157 has side lengths

$$a = \frac{6803298487826435051217540}{411340519227716149383203} \quad \text{and} \quad b = \frac{411340519227716149383203}{21666555693714761309610}.$$

This solution would take a long time to find by a brute force search.

The terminology "congruent" arises from the fact that if n is a congruent number, then there exists a rational number A such that $n - A$, A, and $n + A$ are all rational numbers. Thus n is the common congruence between the rational numbers.

Proposition 13.1.2. *Suppose n is the area of a right triangle with rational side lengths a, b, c, with $a \leq b < c$. Let $A = (c/2)^2$. Then*

$$A - n, \quad A, \quad \textit{and } A + n$$

are all perfect squares of rational numbers.

Proof. We have

$$\begin{aligned} a^2 + b^2 &= c^2 \\ \frac{1}{2}ab &= n \end{aligned}$$

Add or subtract 4 times the second equation to the first to get

$$\begin{aligned} a^2 \pm 2ab + b^2 &= c^2 \pm 4n \\ (a \pm b)^2 &= c^2 \pm 4n \\ \left(\frac{a \pm b}{2}\right)^2 &= \left(\frac{c}{2}\right)^2 \pm n \\ &= A \pm n \end{aligned}$$

□

The following open problem has motivated much of the work in the theory of congruent numbers.

Open Problem 13.1.3. *Give an algorithm which, given n, outputs whether or not n is a congruent number.*

As we will see, this problem is closely related to a problem about elliptic curves.

13.1.1 Congruent Numbers and Elliptic Curves

The following proposition establishes a link between elliptic curves and the congruent number problem. This link connects the congruent number problem with the Birch and Swinnerton-Dyer conjecture.

Proposition 13.1.4. *Let n be a nonzero rational number. There is a bijection between the sets*

$$A = \left\{(x,y,z) \in \mathbb{Q}^3 : \frac{xy}{2} = n, x^2 + y^2 = z^2\right\}$$

and

$$B = \left\{(r,s) \in \mathbb{Q}^2 : s^2 = r^3 - n^2 r, \text{ with } s \neq 0\right\}$$

given by the maps

$$f(x,y,z) = \left(-\frac{ny}{x+z}, 2n^2 x + z\right)$$

and

$$g(r,s) = \left(\frac{n^2 - r^2}{s}, -\frac{2rn}{s}, \frac{n^2 + r^2}{s}\right).$$

Proof. This proposition can be proved using nothing more than lots of elementary algebraic manipulation. By substitution, verify that f and g are well defined, then check that $f(g(r,s)) = (r,s)$. Finally, let K be the field of fractions of $\mathbb{Q}(n,y)[z]/(z^2 - (y^2 + (2n/y)^2))$. If $(x,y,z) \in A$ then $(x,y,z) = (2n/y, y, z)$ and we find, working in K, that $g(f(2n/y,y,z)) = (2n/y, y, z)$.

We illustrate how to verify that the maps are mutually inverse bijections using MAGMA, though one could of course do this by hand. Input the following:

```
R<n,y> := FieldOfFractions(PolynomialRing(Rationals(),2));
S<z>   := PolynomialRing(R);
T<z>   := quo<S | z^2 - (y^2 + (2*n/y)^2)>;
K<z>   := FieldOfFractions(T);
U<r,s> := FieldOfFractions(PolynomialRing(Rationals(),2));
function f(w)
   x,y,z := Explode(w);
   return [-n*y/(x+z), 2*n^2/(x+z)];
end function;
function g(w)
   r,s := Explode(w);
   return [(n^2-r^2)/s, -2*r*n/s, (n^2+r^2)/s];
end function;
```

Then

```
> g(f([2*n/y,y,z]));
[2*n/y, y, z]
> f(g([r,s]));
[n, y]
```

It says (n,y) instead of (r,s) because n and y were the names we first assigned for printing the variables of *the* polynomial ring in two variables (in MAGMA there is only one such ring). □

Corollary 13.1.5. *The nonzero rational number n is a congruent number if and only if the elliptic curve E_n defined by $y^2 = x^3 - n^2x$ has a solution with $y \neq 0$.*

Proof. The number n is a congruent number if and only if the set A from Proposition 13.1.4 is nonempty. By the proposition A is nonempty if and only if B is nonempty, which proves the corollary. $\square$

Example 13.1.6. Let $n = 5$. Then E_n is defined by $y^2 = x^3 - 25x$, and we find by a brute force search the solution $(-4, -6)$. Then

$$g(-4,-6) = \left(\frac{25-16}{-6}, -\frac{-40}{-6}, \frac{25+16}{-6}\right) = \left(-\frac{3}{2}, -\frac{20}{3}, -\frac{41}{6}\right).$$

Multiplying through by -1 yields the side lengths of a rational right triangle with area 5.

Example 13.1.7. Let $n = 1$, so E_1 is defined by $y^2 = x^3 - x$. Since 1 is not a congruent number, the elliptic curve E_1 has no point with $y \neq 0$.

Recall that if A is an abelian group, then the *torsion subgroup* A_{tor} of A is the subgroup of elements of A with finite order.

Proposition 13.1.8. *The torsion subgroup of $E_n(\mathbb{Q})$ has order* 4.

This proposition can be proved by considering natural reduction maps from $E_n(\mathbb{Q})$ to the group of points on the elliptic curve over $\mathbb{F}_p$ defined by $y^2 = x^3 - n^2x$ for many p. For details, see e.g., [**?**, §9] (Neil Koblitz's book "Introduction to Elliptic Curves and Modular Forms").

Recall that the *rank* of an elliptic curve E over $\mathbb{Q}$ is the positive integer r such that $E(\mathbb{Q})/E(\mathbb{Q})_{\mathrm{tor}} \approx \mathbb{Z}^r$. Combining the above corollary and proposition proves the following theorem.

Theorem 13.1.9. *A nonzero rational number n is a congruent number if and only if $E_n(\mathbb{Q})$ has rank ≥ 1.*

The following corollary is not at all obvious from the definition of a congruent number, but it follows immediately from the theorem.

Corollary 13.1.10. *If n is a congruent number, then there are infinitely many right triangles with area $\pm n$.*

In the next section we will associate to any elliptic curve E over $\mathbb{Q}$ a holomorphic function $L(E,s)$ on $\mathbb{C}$. The Birch and Swinnerton-Dyer conjecture predicts that E has positive rank if and only if $L(E,1) = 0$. Using "half integral weight modular forms" and a deep theorem of Waldspurger, Jerrold Tunnell gave a simple criterion for whether or not $L(E_n, 1) = 0$. Thus a proof of the Birch and Swinnerton-Dyer conjecture would also solve Problem 13.1.3.

Theorem 13.1.11 (Tunnell). *If n is an even squarefree integer then $L(E_n, 1) = 0$ if and only if*

$$\#\left\{(a,b,c) : 4a^2 + b^2 + 8c^2 = \frac{n}{2} : c \textit{ is even}\right\}$$

$$= \# \left\{ (a,b,c) : 4a^2 + b^2 + 8c^2 = \frac{n}{2} : c \text{ is odd} \right\}.$$

If n is odd and squarefree then $L(E_n, 1) = 0$ if and only if

$$\# \left\{ (a,b,c) : 2a^2 + b^2 + 8c^2 = n : c \text{ is even} \right\}$$

$$= \# \left\{ (a,b,c) : 2a^2 + b^2 + 8c^2 = n : c \text{ is odd} \right\}.$$

Example 13.1.12. When $n = 6$, we get

$$\#\emptyset = \#\emptyset,$$

and when $n = 1$, we get

$$\#\{(0,1,0)\} \neq \#\emptyset.$$

Partial results are known towards the assertion that $E_n(\mathbb{Q})$ is infinite if and only if $L(E_n, 1) = 0$. The implication "$E_n(\mathbb{Q})$ has positive rank implies that $L(E_n, 1) = 0$" was proved by John Coates and Andrew Wiles. The other implication "$L(E_n, 1) = 0$ implies that $E_n(\mathbb{Q})$ has positive rank" is still unknown today. It was proved in the special case when $L'(E_n, 1) \neq 0$ by Dick Gross and Don Zagier.

13.2 The Birch and Swinnerton-Dyer Conjecture

Let E be the elliptic curve over $\mathbb{Q}$ defined by

$$y^2 = x^3 + ax + b$$

with $a, b \in \mathbb{Z}$ and $\Delta = -16(4a^3 + 27b^2) \neq 0$. For $p \nmid \Delta$, let

$$a_p = p + 1 - \#E(\mathbb{Z}/p\mathbb{Z}).$$

Set

$$L^*(E, s) = \prod_{p \nmid \Delta} \frac{1}{1 - a_p p^{-s} + p^{1-2s}}.$$

Theorem 13.2.1 (Breuil, Conrad, Diamond, Taylor, Wiles). *$L^*(E, s)$ extends to an analytic function on all of $\mathbb{C}$.*

Definition 13.2.2 (Algebraic Rank). The *algebra rank* of E is the unique nonnegative integer r such that $E(\mathbb{Q})/E(\mathbb{Q})_{\mathrm{tor}} \approx \mathbb{Z}^r$.

Definition 13.2.3 (Analytic Rank). The Taylor expansion of $L(E, s)$ at $s = 1$ has the form

$$L^*(E, s) = c(s-1)^r + \text{higher order terms}$$

with $c \neq 0$. This number r is called the *analytic rank* of E.

Conjecture 13.2.4 (Birch and Swinnerton-Dyer). *The analytic and algebraic ranks of E are the same. That is, the order of vanishing of $L^*(E, s)$ at $s = 1$ is the same as the minimal number of generators of $E(\mathbb{Q})/E(\mathbb{Q})_{\mathrm{tor}}$.*

Note that a special case of the conjecture is the assertion that $L(E, 1) = 0$ if and only if $E(\mathbb{Q})$ is infinite. This special case would be enough to give a complete solution to the congruent number problem.

13.2.1 Some of What is Known

Theorem 13.2.5 (Gross, Kolyvagin, Zagier, et al.). *Let E be an elliptic curve. If the analytic rank of E is 0 or 1, then Conjecture 13.2.4 is true.*

It is a folklore conjecture that "most" elliptic curves satisfy the hypothesis of the above theorem, i.e., they have analytic rank 0 or 1. For example, just over 95% of the "first 78198" elliptic curves have analytic rank at most 1. Many mathematicians suspect that the curves with rank bigger than 1 have "density" 0 amongst all elliptic curves. However, in practice it is often the curves of rank bigger than 1 that are interesting.

13.2.2 How to Compute $L(E, s)$ with a Computer

(**Expand this:** "Best Models")
Let E be an elliptic curve over $\mathbb{Q}$, defined by a Weierstrass equation

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6.$$

There are many choices of Weierstrass equations that define an elliptic curve that is "essentially the same" as E. E.g., you found others by completing the square. Among all of these, there is a best possible model, which is the one with smallest discriminant. It can be computed in PARI as follows:

```
? E = ellinit([0,0,0,-43,166]);
? E.disc
%61 = -6815744
? E = ellchangecurve(E,ellglobalred(E)[2])
%62 = [1, -1, 1, -3, 3, ...]
? E.disc
%63 = -1664
```

Thus $y^2 + xy + y = x^3 - x^2 - 3x + 3$ is a "better" model than $y^2 = x^3 - 43x + 166$.

WARNING: Some of the elliptic curves functions in PARI will *LIE* if you give as input an elliptic curve that is defined by a model that isn't the best possible. These devious liars include `elltors, ellap, ellak`, and `elllseries`.
(**Expand this:** "Formula for $L(E, s)$")
As mentioned before, the PARI function `elllseries` can compute $L(E, s)$. I figured out how this function works, and explain it below.

Because E is modular, one can show that we have the following rapidly-converging series expression for $L(E, s)$, for $s > 0$:

$$L(E, s) = N^{-s/2} \cdot (2\pi)^s \cdot \Gamma(s)^{-1} \cdot \sum_{n=1}^{\infty} a_n \cdot (F_n(s-1) - \varepsilon F_n(1-s))$$

where

$$F_n(t) = \Gamma\left(t+1, \frac{2\pi n}{\sqrt{N}}\right) \cdot \left(\frac{\sqrt{N}}{2\pi n}\right)^{t+1}.$$

Here

$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt$$

is the Γ-*function* (e.g., $\Gamma(n) = (n-1)!$), and

$$\Gamma(z, \alpha) = \int_{\alpha}^{\infty} t^{z-1} e^{-t} dt$$

is the *incomplete* Γ-*function.* The number N is called the *conductor* of E and is very similar to the discriminant of E; it is only divisible by primes that divide the best possible discriminant of E. You can compute N using the PARI command `ellglobalred(E)[1]`.

As usual, for $p \nmid \Delta$, we have

$$a_p = p + 1 - \#E(\mathbb{Z}/p\mathbb{Z}),$$

and for $r \geq 2$,

$$a_{p^r} = a_{p^{r-1}} a_p - p a_{p^{r-2}},$$

and $a_{nm} = a_n a_m$ if $\gcd(n, m) = 1$ (I won't define the a_p when $p \mid \Delta$, but it's not difficult.) Finally, ε depends only on E and is either $+1$ or -1. I won't define ε either, but you can compute it in PARI using `ellrootno(E)`.

At $s = 1$, the formula can be massively simplified, and we have

$$L(E, 1) = (1 + \varepsilon) \cdot \sum_{n=1}^{\infty} \frac{a_n}{n} e^{-2\pi n/\sqrt{N}}.$$

This sum converges rapidly, because $e^{-2\pi n/\sqrt{N}} \to 0$ quickly as $n \to \infty$.

13.3 A Rationality Theorem

In the last lecture, I mentioned that it can be surprisingly difficult to say anything precise about $L(E, s)$, even with the above formulas. For example, it is a very deep theorem of Gross and Zagier that for the elliptic curve $y^2 + y = x^3 - 7x + 6$ we have

$$L(E, s) = c(s - 1)^3 + \text{ higher order terms},$$

and nobody has any idea how to prove that there is an elliptic curve with

$$L(E, s) = c(s - 1)^4 + \text{ higher order terms}.$$

Fortunately, it is possible to decide whether or not $L(E, 1) = 0$.

Theorem 13.3.1. *Let $y^2 = x^3 + ax + b$ be an elliptic curve. Let*

$$\Omega_E = 2^n \int_{\gamma}^{\infty} \frac{dx}{\sqrt{x^3 + ax + b}},$$

where γ is the largest real root of $x^3 + ax + b$, and $n = 0$ if $\Delta(E) < 0$, $n = 1$ if $\Delta(E) > 0$. Then

$$\frac{L(E, 1)}{\Omega_E} \in \mathbb{Q},$$

and the denominator is ≤ 24.

In practice, one computes Ω_E using the "Arithmetic-Geometric Mean", *NOT* numerical integration. In PARI, Ω_E is approximated by `E.omega[1]*2^(E.disc>0)`.

Remark 13.3.2. I don't know if the denominator is ever really as big as 24. It would be a fun student project to either find an example, or to understand the proof that the quotient is rational and prove that 24 can be replaced by something smaller.

Example 13.3.3. Let E be the elliptic curve $y^2 = x^3 - 43x + 166$. We compute $L(E,1)$ using the above formula and observe that $L(E,1)/\Omega_E$ appears to be a rational number, as predicted by the theorem.

```
? E = ellinit([0,0,0,-43,166]);
? E = ellchangecurve(E, ellglobalred(E)[2]);
? eps = ellrootno(E)
%77 = 1
? N = ellglobalred(E)[1]
%78 = 26
? L = (1+eps) * sum(n=1,100, ellak(E,n)/n * exp(-2*Pi*n/sqrt(N)))
%79 = 0.6209653495490554663758626727
? Om = E.omega[1]*2^(E.disc>0)
%80 = 4.346757446843388264631038710
? L/Om
%81 = 0.1428571428571428571428571427
? contfrac(L/Om)
%84 = [0, 7]
? 1/7.0
%85 = 0.1428571428571428571428571428
? elltors(E)
%86 = [7, [7], [[1, 0]]]
```

Notice that in this example, $L(E,1)/\Omega_E = 1/7 = 1/\#E(\mathbb{Q})$. This is shadow of a more refined conjecture of Birch and Swinnerton-Dyer.

Example 13.3.4. In this example, we verify that $L(E,1) = 0$ computationally.

```
? E=ellinit([0, 1, 1, -2, 0]);
? L1 = elllseries(E,1)
%4 = -6.881235151133426545894712438 E-29
? Omega = E.omega[1]*2^(E.disc>0)
%5 = 4.980425121710110150642715583
? L1/Omega
%6 = 1.795732353252503036074927634 E-20
```

13.4 Approximating the Rank

Fix an elliptic curve E over $\mathbb{Q}$.

The usual method to *approximate* the rank is to find a series that rapidly converges to $L^{(r)}(E,1)$ for $r = 0,1,2,3,\ldots$, then compute $L(E,1)$, $L'(E,1)$, $L^{(2)}(E,1)$, etc., until one appears to be nonzero. You can read about this method in §2.13 of Cremona's book *Algorithms for Elliptic Curves*. For variety, I will describe a slightly different method that I've played with recently, which uses the formula for $L(E,s)$ from the last lecture, the definition of the derivative, and a little calculus.

Proposition 13.4.1. *Suppose that*

$$L(E,s) = c(s-1)^r + \textit{ higher terms.}$$

Then

$$\lim_{s\to 1}(s-1)\cdot\frac{L'(E,s)}{L(E,s)} = r.$$

Proof. Write

$$L(s) = L(E,s) = c_r(s-1)^r + c_{r+1}(s-1)^{r+1} + \cdots.$$

Then

$$\begin{aligned}\lim_{s\to 1}(s-1)\cdot\frac{L'(s)}{L(s)} &= \lim_{s\to 1}(s-1)\cdot\frac{rc_r(s-1)^{r-1}+(r+1)c_{r+1}(s-1)^r+\cdots}{c_r(s-1)^r+c_{r+1}(s-1)^{r+1}+\cdots}\\ &= r\cdot\lim_{s\to 1}\frac{c_r(s-1)^r+\frac{(r+1)}{r}c_{r+1}(s-1)^{r+1}+\cdots}{c_r(s-1)^r+c_{r+1}(s-1)^{r+1}+\cdots}\\ &= r.\end{aligned}$$

□

Thus the rank r is "just" the limit as $s\to 1$ of a certain (smooth) function. We know this limit is an integer. But, for example, for the curve

$$y^2+xy = x^3-x^2-79x+289$$

nobody has succeeded in proving that this integer limit is 4. (One can prove that the limit is either 2 or 4.)

Using the definition of derivative, we *heuristically* approximate $(s-1)\frac{L'(s)}{L(s)}$ as follows. For $|s-1|$ small, we have

$$\begin{aligned}(s-1)\frac{L'(s)}{L(s)} &= \frac{s-1}{L(s)}\cdot\lim_{h\to 0}\frac{L(s+h)-L(s)}{h}\\ &\approx \frac{s-1}{L(s)}\cdot\frac{L(s+(s-1)^2)-L(s)}{(s-1)^2}\\ &= \frac{L(s^2-s+1)-L(s)}{(s-1)L(s)}\end{aligned}$$

Question 13.4.2. Does

$$\lim_{s\to 1}(s-1)\cdot\frac{L'(s)}{L(s)} = \lim_{s\to 1}\frac{L(s^2-s+1)-L(s)}{(s-1)L(s)}?$$

In any case, we can use this formula in PARI to "approximate" r.

```
? E = ellinit([ 0, 1, 1, -2, 0 ]);
? r(E,s) = L1=elllseries(E,s); L2=elllseries(E,s^2-s+1); (L2-L1)/((s-1)*L1);
? r(E,1.01)
%8 = 2.004135342473941928617680057
? r(E,1.001)
%9 = 2.000431337547225544819319104
\\ One can prove that 2 is the correct limit.
```

Now let's try the mysterious curve $y^2+xy=x^3-x^2-79x+289$ of rank 4:

```
? E=ellinit([ 1,-1,0,-79,289]);
? r(E,1.001)         \\ takes 6 seconds on PIII 1Ghz
%1 = 4.0022223745190856108964406642
? r(E,1.00001)
%2 = 4.000016181256911064613006133
```

It certainly looks like $\lim_{s\to 1} r(s) = 4$. We know for a fact that $\lim_{s\to 1} r(s) \in \mathbb{Z}$, and if only there were a good way to bound the error we could conclude that the limit is 4. But this has stumped people for years, and maybe it is impossible without a very deep result that somehow interprets this limit in a different way. This problem has totally stumped the experts for years. We desperately need a new idea!!

If one of you wants to do a reading or research project on this problem in the next year or two, let me know. One could draw pictures of $L^{(3)}(E,s)$ or investigate the analogous problem for other more accessible L-series.

```
? E=ellinit([0,0,1,-7,6]);
? r(E,s) = L1=elllseries(E,s); L2=elllseries(E,s^2-s+1); (L2-L1)/((s-1)*L1);
? r(E,1.001)
%2 = 3.0011441049856192065044448552
```

Part III

Computing

14
Computing With PARI/GP

"The object of numerical computation is theoretical advance."
– *Bryan Birch describing A. O. L. Atkin.*

14.1 Introduction

Much progress in number theory has been driven by attempts to prove conjectures. It's reasonably easy to play around with integers, see a pattern, and make a conjecture. Frequently proving the conjecture is *extremely difficult.* In this direction, computers help us to

- find more conjectures
- disprove conjectures
- increase our confidence in a conjecture

They also frequently help to solve a specific problem. For example, the following problem would be hopelessly tedious by hand. Here's an example of such a problem:

> Find all integer $n < 100$ that are the area of a right triangle with integer side lengths.[1]

This problem can be solved by a combination of very deep theorems, a few big computer computations, and a little luck.

[1] We will discuss the "The Congruent Number Problem" in more depth later in this course.

14.1.1 Some Assertions About Primes

A computer can quickly "convince" you that many assertions about prime numbers are true. Here are three.

- *The polynomial $x^2 + 1$ takes on infinitely many prime values.*
  Let
  $$f(n) = \{x : x < n : x \text{ and } x^2 + 1 \text{ is prime }\}.$$
  With a computer, we quickly find that
  $$f(10^2) = 19, \quad f(10^3) = 112, \quad f(10^4) = 841, \quad f(10^5) = 6656.$$
  Surely $f(n)$ is unbounded! The PARI code to compute $f(n)$ is very simple:

```
? f(n) = s=0; for(x=1,n,if(isprime(x^2+1),s++)); s
? f(100)
%1 = 19
? f(1000)
%2 = 112
? f(10000)
%3 = 841
? f(100000)
%4 = 6656
```

- *Every even integer $n > 2$ is a sum of two primes.*
  With a computer we find that this seems true

n	p	q
4	2	2
6	3	3
8	3	5
10	3	7
12	5	7

  ... and much further. In practice, it's easy to write an even number as a sum of two primes. Why should there be any weird even numbers out there for which this can't be done? PARI code to find p and q:

```
? gb(n) = forprime(p=2,n,if(isprime(n-p),return([p,n-p])));
? gb(4)
%7 = [2, 2]
? gb(6)
%8 = [3, 3]
? gb(100)
%9 = [3, 97]
? gb(1000)
%10 = [3, 997]
? gb(570)          \\ takes no time at all!
%11 = [7, 563]
```

- *There are infinitely many primes p such that $p+2$ is also prime.* Let $t(n) = \#\{p : p \leq n \text{ and } p+2 \text{ is prime }\}$. Using a computer we quickly find that

  $$t(10^2) = 8, \quad t(10^3) = 35, \quad t(10^4) = 205, \quad t(10^5) = 1024.$$

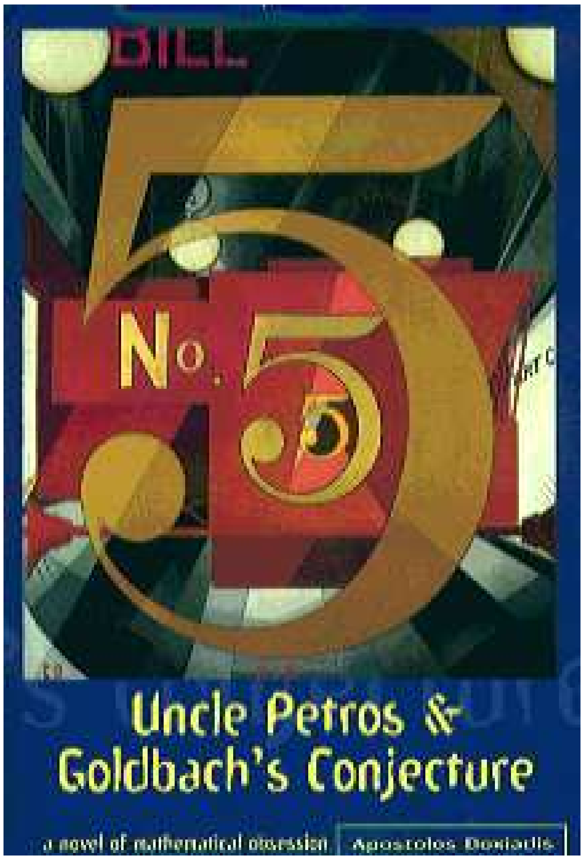
  The PARI code to compute $t(n)$ is very simple:

```
? t(n) = s=0; forprime(p=2,n,if(isprime(p+2),s++)); s
? t(10^2)
%12 = 8
? t(10^3)
%13 = 35
? t(10^4)
%14 = 205
? t(10^5)
%15 = 1224
```

  Surely $t(n)$ keeps getting bigger!!

As it turns out, these three assertions are *all* OLD famous extremely difficult <u>unsolved</u> problems! Anyone who proves one of them will be very famous.

Assertion 2 is called "The Goldbach Conjecture"; Goldbach reformulated it in a letter to Euler in 1742. It's featured in the following recent novel:



The publisher of that novel offers a MILLION dollar prize for the solution to the Goldbach conjecture:

```
http://www.faber.co.uk/faber/million_dollar.asp?PGE=&ORD=faber&TAG=&CID=
```

The Goldbach conjecture is true for all $n < 4 \cdot 10^{14}$, see

```
http://www.informatik.uni-giessen.de/staff/richstein/ca/Goldbach.html
```

Assertion 3 is the "Twin Primes Conjecture". According to

```
http://perso.wanadoo.fr/yves.gallot/primes/chrrcds.html#twin
```

on May 17, 2001, David Underbakke and Phil Carmody discovered a 32220 digits twin primes record with a set of different programs: $318032361 \cdot 2^{107001} \pm 1$. This is the current "world record".

With a computer, even if you can't solve one of these "Grand Challenge" problems, at least you can perhaps work very hard and prove it for more cases than anybody before you, especially since computers keep getting more powerful. This can be very fun, especially as you search for a more efficient algorithm to extend the computations.

14.1.2 Some Tools for Computing

Calculator: A TI-89 can deal with integers with 1000s of digits, factor, and do most basic number theory. I am not aware if anyone has programmed basic "elliptic curve" computations into this calculator, but it could be done.

Mathematica and Maple: Both are commercial, but they are very powerful, can draw pretty pictures, and there are elliptic curve packages available for each (`apecs` for Maple, and something by Silverman for Mathematica).

PARI: Free, open source, excellent for our course, simple, runs on Macs, MS Windows, Linux, etc.

MAGMA: Huge, non-free but nonprofit, what I usually use for my research. I can legally give you a Linux executable if you are registered for 124.

My Wristwatch: Perhaps the only wristwatch in the world that can factor your social security number? `:-)`

14.1.3 Getting Started with PARI

(**Expand this:** "Documentation")
The documentation for PARI is available at

`http://modular.fas.harvard.edu/docs/`

Some PARI documentation:

1. **Installation Guide:** Help for setting up PARI on a UNIX computer.
2. **Tutorial:** 42-page tutorial that starts with `2 + 2`.
3. **User's Guide:** 226-page reference manual; describes every function
4. **Reference Card:** hard to print, so I printed it for you (handout)

(**Expand this:** "A Short Tour")

```
$ gp
Appele avec : /usr/local/bin/gp -s 10000000 -p 500000 -emacs

                 GP/PARI CALCULATOR Version 2.1.1 (released)
               i686 running linux (ix86 kernel) 32-bit version
               (readline v4.2 enabled, extended help available)

                     Copyright (C) 2000 The PARI Group

PARI/GP is free software, covered by the GNU General Public License, and
comes WITHOUT ANY WARRANTY WHATSOEVER.

Type ? for help, \q to quit.
Type ?12 for how to get moral (and possibly technical) support.
```

```
   realprecision = 28 significant digits
   seriesprecision = 16 significant terms
   format = g0.28

parisize = 10000000, primelimit = 500000
? \\ this is a comment
? x = 571438063;
? print(x)
571438063
? x^2+17
%2 = 326541459845191986
? factor(x)
%3 =
[7 1]

[81634009 1]
? gcd(x,56)
%5 = 7
? x^20
%6 = 13784255037665854930357784067541250773222915495828020913935
8450113971943932613097560462268162512901194466231159983662241797
60816483100648674388195744425584150472890085928660801
```

(**Expand this:** “Help in PARI”)

```
? ?
Help topics:
  0: list of user-defined identifiers (variable, alias, function)
  1: Standard monadic or dyadic OPERATORS
  2: CONVERSIONS and similar elementary functions
  3: TRANSCENDENTAL functions
  4: NUMBER THEORETICAL functions
  5: Functions related to ELLIPTIC CURVES
  6: Functions related to general NUMBER FIELDS
  7: POLYNOMIALS and power series
  8: Vectors, matrices, LINEAR ALGEBRA and sets
  9: SUMS, products, integrals and similar functions
 10: GRAPHIC functions
 11: PROGRAMMING under GP
 12: The PARI community

Further help (list of relevant functions): ?n (1<=n<=11).
Also:
  ? functionname (short on-line help)
  ?\             (keyboard shortcuts)
  ?.             (member functions)
Extended help looks available:
  ??             (opens the full user's manual in a dvi previewer)
```

```
  ??  tutorial   (same with the GP tutorial)
  ??  refcard    (same with the GP reference card)

  ??  keyword    (long help text about "keyword" from the user's manual)
  ??? keyword    (a propos: list of related functions).
? ?4

addprimes     bestappr      bezout        bezoutres     bigomega
binomial      chinese       content       contfrac      contfracpnqn
core          coredisc      dirdiv        direuler      dirmul
divisors      eulerphi      factor        factorback    factorcantor
factorff      factorial     factorint     factormod     ffinit
fibonacci     gcd           hilbert       isfundamental isprime
ispseudoprime issquare      issquarefree  kronecker     lcm
moebius       nextprime     numdiv        omega         precprime
prime         primes        qfbclassno    qfbcompraw    qfbhclassno
qfbnucomp     qfbnupow      qfbpowraw     qfbprimeform  qfbred
quadclassunit quaddisc      quadgen       quadhilbert   quadpoly
quadray       quadregulator quadunit      removeprimes  sigma
sqrtint       znlog         znorder       znprimroot    znstar

? ?gcd
gcd(x,y,{flag=0}): greatest common divisor of x and y. flag is optional, and
can be 0: default, 1: use the modular gcd algorithm (x and y must be
polynomials), 2 use the subresultant algorithm (x and y must be polynomials).

? ??gcd
\\ if set up correctly, brings up the typeset subsection from the manual on gcd
```

14.2 Pari Programming

14.2.1 Beyond One Liners

In today's relaxing but decidely non-mathematical lecture, you will learn a few new PARI programming commands. Feel free to try out variations of the examples below (especially because there is no homework due this coming Wednesday). Also, given that you know PARI fairly well by now, ask me questions during today's lecture!
(**Expand this:** "Reading Files")
The `\r` command allows you to read in a file.

Example 14.2.1. Create a file `pm.gp` that contains the following lines

```
{powermod(a, p, n) =
    return (lift(Mod(a,p)^n));}
```

Now use `\r` to load this little program into PARI:

```
> ?powermod
  ***   powermod: unknown identifier.
```

```
> \rpm                \\ \rpm.gp would do the same thing
? ?powermod
powermod(a, p, n) = return(lift(Mod(a,p)^n));
? powermod(2,101,7)
%1 = 27
```

If we change pm.gp, just type \r to reload it (omitting the file name reloads the last file loaded). For example, suppose we change `return (lift(Mod(a,p)^n))` in pm.gp to `return (lift(Mod(a,p)^n)-p)`. Then

```
? \r
? powermod(2,101,7)
%2 = -74
```

(**Expand this:** "Arguments")
PARI functions can have several arguments. For example,

```
{add(a, b, c)=
   return (a + b + c);}
? add(1,2,3)
%3 = 6
```

If you leave off arguments, they are set equal to 0.

```
? add(1,2)
%4 = 3
```

If you want the left-off arguments to default to something else, include that information in the declaration of the function:

```
{add(a, b=-1, c=2)=
   return (a + b + c);}
? add(1,2)
%6 = 5
? add(1)
%7 = 2
? add(1,2,3)
%8 = 6
```

(**Expand this:** "Local Variables Done Right")
Amidst the haste of a previous lecture, I mentioned that an unused argument can be used as a poor man's local variable. The following example illustrates the right way to declare local variables in PARI.

Example 14.2.2. The function `verybad` below sums the integers $1, 2, \ldots n$ whilst wreaking havoc on the variable `i`.

```
{verybad(n)=
   i=0;
   for(j=1,n, i=i+j);
   return(i);}
? verybad(3)
%9 = 6
? i=4;
? verybad(3);
```

```
? i
%13 = 6                    \\ ouch!! what have you done to my eye!
```

The function `poormans` is better, but it uses a cheap hack to simulate a local variable.

```
{poormans(n,  i=0)=
   for(j=1,n, i=i+j);
   return(i);}
? i=4;
? poormans(3)
%16 = 6
? i
%17 = 4              \\ good
```

The following function is the best, because i is local and it's clearly declared as such.

```
{best(n)=
   local(i);
   i=0;  for(j=1,n, i=i+j);
   return(i);}
? i=4;
? best(3)
%18 = 6
? i
%19 = 4
```

(**Expand this:** "Making Your Program Listen")
The `input` command reads a PARI expression from the keyboard. The expression is evaluated and the result returned to your program. This behavior is at first disconcerting if, like me, you naively expect `input` to return a string. Here are some examples to illustrate the `input` command:

```
? ?input
input(): read an expression from the input file or standard input.
? s = input();
1+1
? s                  \\ s is not the string "1+1", as you might expect
%24 = 2
? s=input()
hi there
%25 = hithere
? type(s)            \\ PARI views s as a polynomial in the variable hithere
%26 = "t_POL"
? s=input()
"hi there"
%27 = "hi there"
? type(s)            \\ now it's a string
%28 = "t_STR"
```

(**Expand this:** "Writing to Files")
Use the `write` command:

```
? ?write
write(filename,a): write the string expression a to filename.
? write("testfile", "Hello Kitty!")
```

The `write` command above appended the line "Hello Kitty!" to the last line of `testfile`. This is useful if, e.g., you want to save key bits of work during a session or in a function. There is also a **logging facility** in PARI, which records most of what you type and PARI outputs to the file `pari.log`.

```
? \l
   log = 1 (on)
? 2+2
%29 = 4
? \l
   log = 0 (off)
   [logfile was "pari.log"]
```

14.2.2 Coming Attractions

The rest of this course is about continued fractions, quadratic forms, and elliptic curves. The following illustrates some relevant PARI commands which will help us to explore these mathematical objects.

```
? ?contfrac
contfrac(x,{b},{lmax}): continued fraction expansion of x ...
? contfrac(7/9)
%30 = [0, 1, 3, 2]
? contfrac(sqrt(2))
%31 = [1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...]
? ?qfbclassno
qfbclassno(x,{flag=0}): class number of discriminant x using Shanks's
method by default. If (optional) flag is set to 1, use Euler products.
? qfbclassno(-15,1)  \\ ALWAYS use flag=1, since ``the authors were too
%32 = 2              \\ lazy to implement Shanks' method completely...''
? E=ellinit([0,1,1,-2,0]);
? P=[0,0];
? elladd(E,P,P)
%36 = [3, 5]
? elladd(E,P,[3,5])
%37 = [-11/9, 28/27]
? a=-11/9;b=28/27;          \\ this is an ``amazing'' point on the curve.
? b^2+b == a^3+a^2-2*a
%38 = 1
```

14.3 Computing with Elliptic Curves

14.3.1 Initializing Elliptic Curves

We are concerned primarily with elliptic curves E given by an equation of the form

$$y^2 = x^3 + ax + b$$

with a and b either rational numbers or elements of a finite field $\mathbb{Z}/p\mathbb{Z}$. If a and b are in $\mathbb{Q}$, we initialize E in PARI using the following command:

```
? E = ellinit([0,0,0,a,b]);
```

If you wish to view a and b as element of $\mathbb{Z}/p\mathbb{Z}$, initialize E as follows:

```
? E = ellinit([0,0,0,a,b]*Mod(1,p));
```

If $\Delta = -16(4a^3+27b^2) = 0$ then `ellinit` will complain; otherwise, `ellinit` returns a 19-component vector of information about E. You can access some of this information using the dot notation, as shown below.

```
? E = ellinit([0,0,0,1,1]);
? E.a4
%11 = 1
? E.a6
%12 = 1
? E.disc
%13 = -496
? E.j
%14 = 6912/31
? E5 = ellinit([0,0,0,1,1]*Mod(1,5));
? E5.disc
%15 = Mod(4, 5)
? E5.j
%16 = Mod(2, 5)
```

Here `E.j` is the *j-invariant* of E. It is equal to $\frac{2^8 3^3 a^3}{4a^3+27b^2}$, and has some remarkable properties that I probably won't tell you about.

Most elliptic curves functions in PARI take as their first argument the output of `ellinit`. For example, the function `ellisoncurve(E,P)` takes the output of `ellinit` as its first argument and a point `P=[x,y]`, and returns `1` if `P` lies on `E` and `0` otherwise.

```
? P = [0,1]
? ellisoncurve(E, P)
%17 = 1
? P5 = [0,1]*Mod(1,5)
? ellisoncurve(E5, P)
%18 = 1
```

14.3.2 Computing in The Group

The following functions implement some basic arithmetic in the group of points on an elliptic curve: `elladd`, `ellpow`, and `ellorder`. The `elladd` function simply adds together two points using the group law. Warning: PARI does *not* check that the two points are on the curve.

```
? P = [0,1]
%2 = [0, 1]
? elladd(E,P,P)
%3 = [1/4, -9/8]
? elladd(E,P,[1,0])    \\ nonsense, since [1,0] isn't even on E!!!
%4 = [0, -1]
? elladd(E5,P5,P5)
%12 = [Mod(4, 5), Mod(2, 5)]
? [1/4,-9/8]*Mod(1,5)
%13 = [Mod(4, 5), Mod(2, 5)]
```

The `ellpow` function computes $nP = P + P + \cdots + P$ (n summands).

```
? ellpow(E,P,2)
%5 = [1/4, -9/8]
? ellpow(E,P,3)
%6 = [72, 611]
? ellpow(E,P,15)
%7 = [26449452347718826171173662182327682047670541792/9466094804586385762312509661837302961354550401,
4660645813671121765025590267647300672252945873586541077711389394563791/920992883734992462745141522111225908861976098219465616585649245395649]
```

14.3.3 The Generating Function $L(E,s)$

Suppose E is an elliptic curve over $\mathbb{Q}$ defined by an equation $y^2 = x^3+ax+b$. Then for every prime p that does not divide $\Delta = -16(4a^3 + 27b^2)$, the same equation defines an elliptic curve over the finite field $\mathbb{Z}/p\mathbb{Z}$. As you will discover in problem 3 of homework 9, it can be exciting to consider the package of numbers $\#E(\mathbb{Z}/p\mathbb{Z})$ of points on E over all finite fields. The function `ellap` computes

$$a_p(E) = p + 1 - \#E(\mathbb{Z}/p\mathbb{Z}).$$

```
? E = ellinit([0,0,0,1,1]);
? ellap(E,5)
%19 = -3          \\ this should be 5+1 - #points
? E5 = ellinit([0,0,0,1,1]*Mod(1,5));
? for(x=0,4, for(y=0,4, if(ellisoncurve(E5,[x,y]),print([x,y]))))
[0, 1]
[0, 4]
[2, 1]
[2, 4]
[3, 1]
[3, 4]
[4, 2]
[4, 3]
? 5+1 - 9          \\ 8 points above, plus the point at infinity
%22 = -3
```

There is a natural way to extend the definition of a_p to define integers a_n for every integer n. For example, if a_p and a_q are defined as above and p

and q are distinct primes, then $a_{pq} = a_p a_q$. Today I won't tell you how to define the a_p when, e.g., $p \mid \Delta$. However, you can compute the numbers a_n quickly in PARI using the function `ellan`, which computes the first few a_n.

```
? ellan(E,15)
%24 = [1, 0, 0, 0, -3, 0, 3, 0, -3, 0, -2, 0, -4, 0, 0]
```

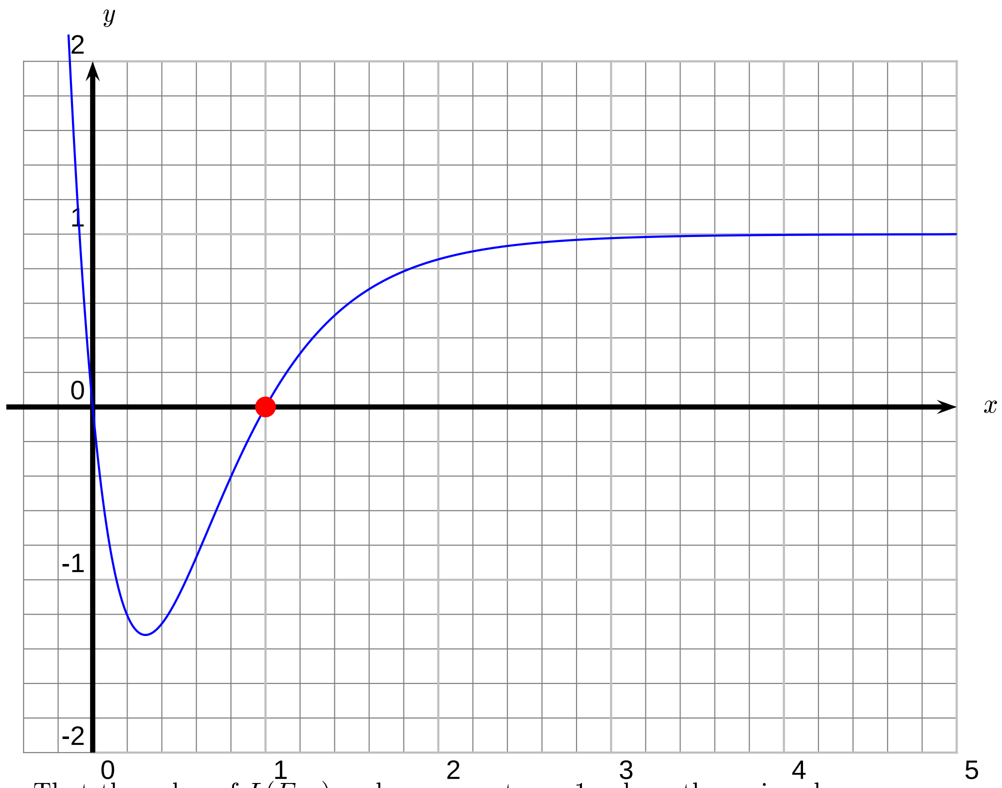
This output means that $a_1 = 1$, $a_2 = a_3 = a_4 = 0$, $a_5 = -3$, $a_6 = 0$, and so on.

When confronted by a mysterious list of numbers, it is a "reflex action" for a mathematician to package them together in a generating function, and see if anything neat happens. It turns out that for the above numbers, a good way to do this is as follows. Define

$$L(E,s) = \sum_{n=1}^{\infty} a_n n^{-s}.$$

This might remind you of Riemann's ζ-function, which is the function you get if you make the simplest generating function $\sum_{n=1}^{\infty} n^{-s}$ of this form.

Using `elllseries(E,s,1)` I drew a graph of $L(E,s)$ for $y^2 = x^3 + x + 1$.



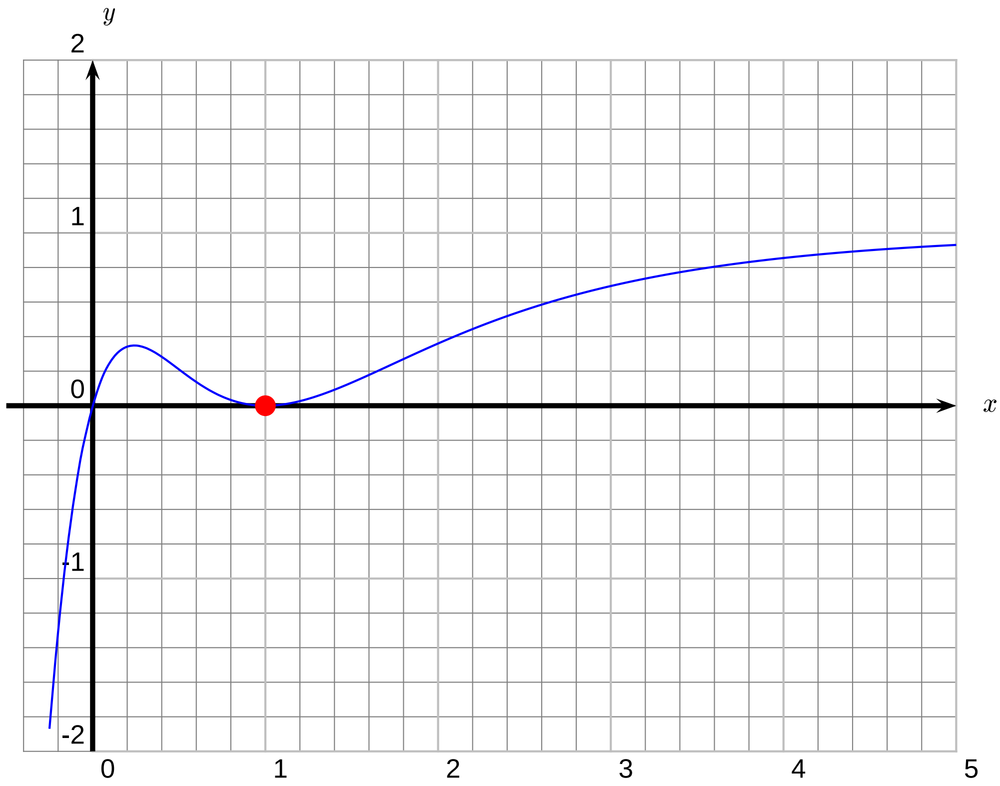
That the value of $L(E,s)$ makes sense at $s = 1$, where the series above doesn't obviously converge, follows from the nontrivial fact that the function

$$f(z) = \sum_{n=1}^{\infty} a_n e^{2\pi i n z}$$

is a *modular form*. Also, keep your eyes on the dot; it plays a central roll in the Birch and Swinnerton-Dyer conjecture, which asserts that $L(E, 1) = 0$ if and only if the group $E(\mathbb{Q})$ is infinite.
(**Expand this:** "A Curve of Rank Two")
Let E be the simplest rank 2 curve:

$$y^2 + y = x^3 + x^2 - 2x.$$

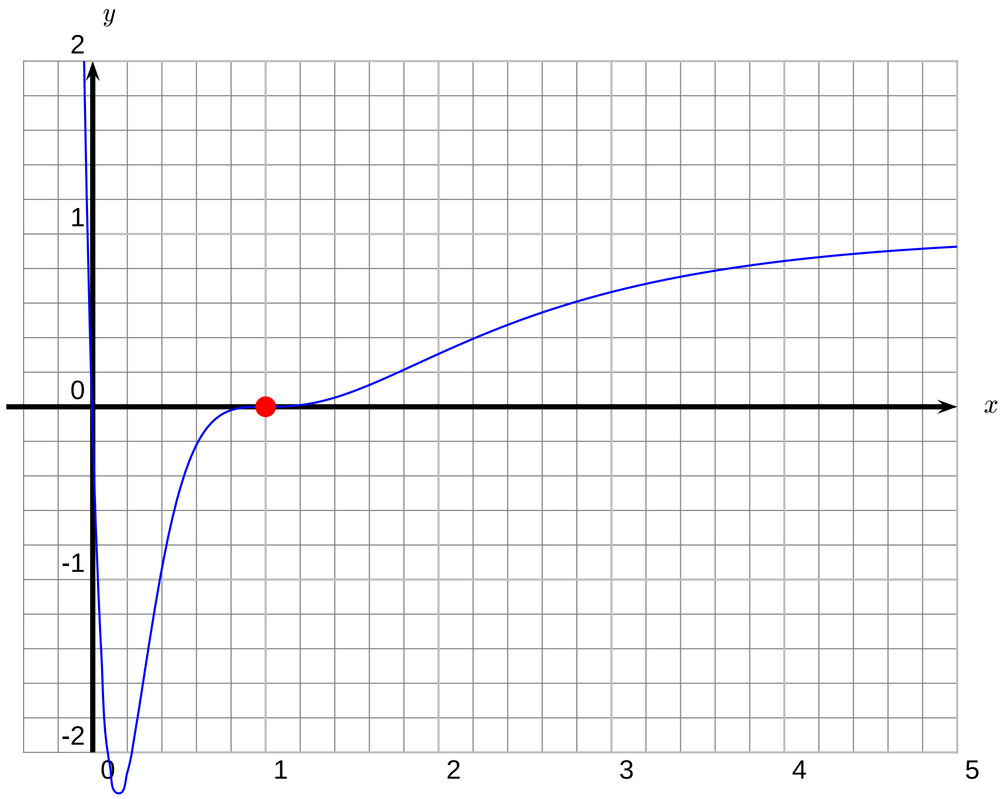
The discriminant is 389.



(**Expand this:** "A Curve of Rank Three")
Let E be the simplest rank 3 curve:

$$y^2 + y = x^3 - 7x + 6.$$

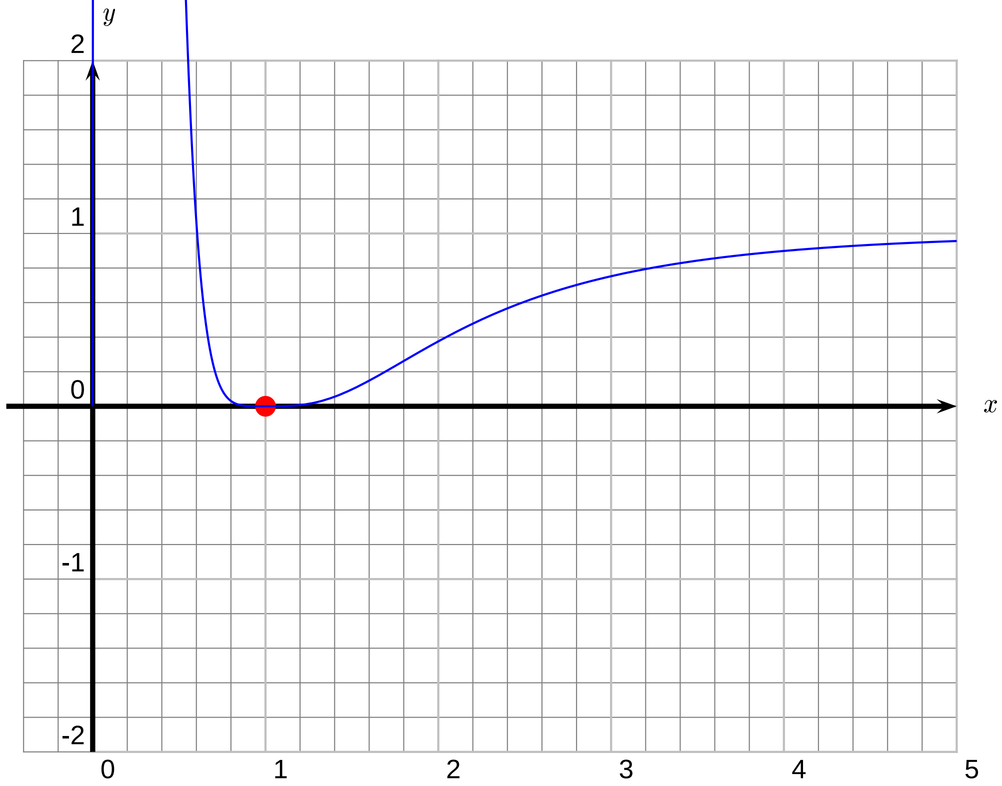
The discriminant is 5077.



(**Expand this:** "A Curve of Rank Four")
Let E be the simplest *known* rank 4 curve:

$$y^2 + xy = x^3 - x^2 - 79x + 289$$

The conductor is $2 \cdot 117223$.



14.3.4 Other Functions and Programs

You can see a complete list of elliptic-curves functions by typing `?5`:

```
? ?5
elladd          ellak           ellan           ellap
ellbil          ellchangecurve  ellchangepoint  elleisnum
elleta          ellglobalred    ellheight       ellheightmatrix
ellinit         ellisoncurve    ellj            elllocalred
elllseries      ellorder        ellordinate     ellpointtoz
ellpow          ellrootno       ellsigma        ellsub
elltaniyama     elltors         ellwp           ellzeta         ellztopoint
```

I have only described a small subset of these. To understand many of them, you must first learn how to view an elliptic curve as a "donut", that is, as quotient of the complex numbers by a *lattice*, and also as a quotient of the upper half plane.

There is a Maple package called APECS for computing with elliptic curves, which is more sophisticated than PARI in certain ways, especially in connection with algorithms that involve lots of commutative algebra. MAGMA also offers sophisticated features for computing with elliptic curves, which are built in to the standard distribution. I will give a demonstrations of MAGMA in the Basic Notions seminar at 3pm on Monday, December 3 in SC 507. There is also a C++ library called LiDIA that has libraries with some powerful elliptic curves features.

15
Programming MAGMA

MAGMA is an excellent tool for computations of an algebraic nature, e.g., finite group theory, combinatorics, computations with basic number theoretic objects, and working with elliptic curves. However, my TI-89 hand calculator is better at computing integrals than MAGMA.

MAGMA has good support for developing large programs and combining code from many projects together. MAGMA's rigorous approach to computer algebra avoids much of the ambiguity that affects some other systems, and forces the user to produce more meaningful code that is easier to read and quicker. MAGMA also has highly optimized support for linear algebra over the rational numbers and $\mathbb{Z}/p$.

This chapter focuses on what MAGMA is and how to use it as a tool to accomplish more than a few quick computations in the shell. We do not dwell on specific MAGMA packages or functions.

15.1 Documentation

Thousands of pages have been written about MAGMA:

`http://magma.maths.usyd.edu.au/magma/htmlhelp/doc.htm`

Invest an hour and read the 12-page *First Steps in MAGMA*, then skim through the 884-page *Introduction*.

Instead of using the help system that is built into the MAGMA shell, I use the HTML reference manual. To look up a command, go to the index for the first letter of the command, then use your browser's find function to find the command, then click on the link. This will lead you to the help for the command, and you can easily navigate up in order to get information about how that command fits in with other commands.

You can also get documentation about the ways to call a command by typing its name, for example:

```
> PolynomialRing;
Intrinsic 'PolynomialRing'
Signatures:
(<RngInvar> R) -> RngMPol
The generic polynomial ring in which the elements of R lie
(<Rng> R) -> RngUPol
[
Global: BoolElt
]
Create the univariate polynomial ring over R
[... etc. for a page]
```

Notice that the behavior of `PolynomialRing` depends on the type of argument you give it.

15.2 Getting Comfortable

Once installed, if you run MAGMA you get a shell in which you can type commands. Without some customization, you will probably soon become impatient with the shell. You should do the following:

1. Create a directory, `magma` say, in which you will store MAGMA files.
2. Create a startup file, e.g., `startup.m`, which will be executed when you start MAGMA. (See Section 15.2.1.)
3. Create a `spec` file, which lists the filenames of code that you want to *attach* attach to MAGMA. (See Section 15.2.2.)
4. Learn to log your sessions to a file, and save and restore them. (See Section 15.2.3.)
5. If you want to use the MAGMA shell under another editor like the emacs shell window, type the command `SetLineEditor(false);` into MAGMA.

15.2.1 Startup File

MAGMA assumes nothing. Some new MAGMA users are frightened when they do the following:

```
[joesixpack@couch]# magma
Magma V2.9-11   [...]
> f := x^2 + 1;
>> f := x^2 + 1;
        ^
User error: Identifier 'x' has not been declared or assigned
```

Like in many strongly typed languages, it is necessary to define `x` first.

```
> R<x> := PolynomialRing(RationalField());
> R;
Univariate Polynomial Ring in x over Rational Field
> f := x^2 + 1;
```

Next, you might be put off by having to type huge words like

`PolynomialRing` and `RationalField`,

but this source of frustration can also be easily circumvented:

```
> poly := PolynomialRing;
> Q := RationalField();
> R<x> := poly(Q);
> R;
Univariate Polynomial Ring in x over Rational Field
```

After typing those first two lines, for the rest of the session you can type `poly` wherever you would type `PolynomialRing` and `Q` where you would have typed `RationalField()`. To keep all of these customization from session to session, create a startup file. For example, make a file `startup.m` that contains the following lines:

```
poly := PolynomialRing;
Q := RationalField();
Z := IntegerRing();
R := RealField();
R<x> := poly(Q);
charpoly := CharacteristicPolynomial;
```

Then set the environment variable `MAGMA_STARTUP_FILE` to `startup.m` (with proper path). Henceforth whenever you start MAGMA, `Q` will be the rationals, and `charpoly` will be the same as `CharacteristicPolynomial`. (Note: When you use the MAGMA shell, if you press tab, MAGMA will do auto-completion.)

15.2.2 Spec File

As we will see in Section 15.3, the MAGMA programs you write are stored in files that you attach to MAGMA.

It is tedious attaching a file to MAGMA each time you start MAGMA, so if you set the environment variable `MAGMA_USER_SPEC` to `$HOME/magma/spec` and list the filenames to attach in `spec`, they will automatically be attached when you start MAGMA.

15.2.3 Logging, Saving, and Restoring

It's frustrating to do something using MAGMA, only to lose the steps of the computation because they've scrolled off the screen. Use the command `SetLogFile("logfile")`, which takes one argument, the name of a file, and appends a log of the current magma session to that file. Type `UnsetLogFile()` to turn off logging.

If you're in the middle of a MAGMA session, and would like to leave and come back to it later, type `save "session"` then quit MAGMA. After you restart MAGMA, type `restore "session"`. (Warning: If you install a new version of MAGMA, the session files you used under the previous version of MAGMA might not load anymore.)

15.3 Programming

The MAGMA programming language resembles many standard procedural languages. Code is divided into files, and the code in files are divided into "functions", "procedures", and "intrinsics". Functions have arguments and return a single value, like in many other languages. A procedure is exactly the same as a function, but it doesn't return a value. Whereas functions and procedures have file scope, intrinsics are exported to the MAGMA shell, and are indistiguishable to the user from any of the other built in MAGMA commands. When you write an intrinsic you extend the MAGMA shell.

Let's extend MAGMA by adding a command called `MySqrt` that computes a square root of any square in $\mathbb{Z}/p$. (This is for fun, since the built in command `IsSquare` already does this.) First create a file called `mysqrt.m` that contains the following lines.

```
function alg3(a)
   assert Type(a) eq RngIntResElt;
   p := Modulus(Parent(a));
   assert p mod 4 eq 3;
   return a^((p+1) div 4);
end function;

function alg1(a)
   assert Type(a) eq RngIntResElt;
   p := Modulus(Parent(a));
   assert p mod 4 eq 1;
   F := Parent(a);
   R<x> := PolynomialRing(F);
   Q<x> := quo<R|x^2-a>;
   while true do
      z := Random(F);
      w := (1+z*x)^((p-1) div 2);
      if Coefficient(w,0) eq 0 then
         return 1/Coefficient(w,1);
      end if;
   end while;
end function;

intrinsic MySqrt(a::RngIntResElt) -> RngIntResElt
{The square root of a.  We assume that a has a square root
 and that a is an element of Z/p with p prime.}
   p := Modulus(Parent(a));
```

```
   if p eq 2 then
      return a;
   end if;
   if a eq 0 then
      return a;
   end if;
   require IsPrime(p) :
              "The modulus of argument 1 must be prime.";
   require KroneckerSymbol(Integers()!a,p) eq 1 :
              "Argument 1 must have a square root.";
   if p mod 4 eq 3 then
      return alg3(a);
   else
      return alg1(a);
   end if;
end intrinsic;
```

There are `assert` statements in the functions because MAGMA does no type checking for arguments to functions, so we have to fake it. Incidentally, we discover that elements of $\mathbb{Z}/p$ are of type `RngIntResElt` by creating an element in the shell and asking for its type:

```
> Type(ResidueClassRing(5)!1);
RngIntResElt
```

We don't pass the modulus p to `alg3` and `alg1`, because a is an element of $\mathbb{Z}/p$ so the function only needs to know a, since a knows $\mathbb{Z}/p$, in the sense that the `Parent` of a is $\mathbb{Z}/p$. To discover the `Modulus` command, I looked up `ResidueClassRing` in the MAGMA HTML documentation, then looked at nearby commands until I saw one called `Modulus`.

The `assert p mod 4 eq 3` line illustrates a healthy level of paranoia. The return line does the square root computation then returns it.

The function `alg1` computes the square root in the case $p \equiv 1 \pmod 4$. After the usual type checking assertion, we create the quotient ring

$$R = (\mathbb{Z}/p)[x]/(x^2 - a).$$

We then raise random elements of the form $1 + zx$ to the power $(p-1)/2$ until finding one of the form vx. The answer is then $1/v$.

Everything is tied together and exported to MAGMA in the intrinsic, which is the last part of the file. The declaration of the intrinsic gives the type of the arguments (multiple arguments are allowed), the return type (multiple return values are allowed), and a *mandatory comment* which must be given in braces. Note that non-intrinsic comments in MAGMA use the usual C++ syntax (/* and */ and //.) After the comment we use `if` statements to treat two special cases. The `require` statement makes certain assertions about the input; if they fail the corresponding error message is printed and execution stops.

To make use of our new function, add the line `mysqrt.m` to your `spec` file. When you start MAGMA the command `MySqrt` will automatically be available. Alternatively, instead of adding `mysqrt.m` to your `spec` file, you can type `Attach("mysqrt.m")` in MAGMA, but this only survives until you exit MAGMA.

If while running MAGMA you edit the file `mysqrt.m`, the changes automatically take affect. There is no need to restart MAGMA.

Here's an example session:

```
> Attach("mysqrt.m");
> R := ResidueClassRing(37);
>> MySqrt(R!13);
         ^
Runtime error in 'MySqrt': Argument 1 must have a square root.
> MySqrt(R!11);
14
> R!14^2;
11
> MySqrt(R!11);
23
> R!23^2;
11
> R := ResidueClassRing(31);
> MySqrt(R!7);
10
> R!10^2;
7
> MySqrt(R!11);
>> MySqrt(R!11);
         ^
Runtime error in 'MySqrt': Argument 1 must have a square root.
> MySqrt(R!19);
9
> R!9^2;
19
```

We can also try large primes to see if the algorithm is at all efficient.

```
> p := NextPrime(0495959487929484949494928292049494891 3);
> p mod 4;
1
> R := ResidueClassRing(p);
> time MySqrt(R!5);             // time times the command
45065146537549164856318874663544056 3
Time: 0.150
> $1^2;                         // $1 means the last output.
5
```

15.4 Weaknesses

I think the biggest weakness of MAGMA is that it is a strongly typed language with some object orient constructions, but it is almost impossible for the user to define new data types. The types that come with MAGMA are extensive, including rings, fields, vector spaces, modules, elliptic curves, etc., but they don't begin to encompass all of mathematics. Users cannot derive from built in types, though the `declare attributes` construction provides a very weak form of derivation. There are two unsatisfactory ways to add a new type to MAGMA: one is to become friends with the developers of MAGMA and ask them to add a new type to the kernel; the other is to find out about secret "Hack objects", which are blank types with names like "Xxx" that are compiled into MAGMA with no predefined functionality.

A mild frustration with MAGMA is that it is impossible to read in MAGMA code from a running MAGMA program. This makes saving and later reloading MAGMA variables very tedious. The best way to get around this is to use another language, such as python or perl, to set up computations, then run MAGMA via a system call.

The core of MAGMA is closed source, which is undesirable in a tool that one wants to use to prove rigorous mathematical results. This can be frustrating, but fortunately MAGMA is nonprofit and the developers of MAGMA are friendly and responsive, so it is possible to find out what MAGMA is really doing in any particular situation by asking them.

MAGMA is a huge piece of software; for example, there are nearly 250000 lines of package code included with MAGMA, and the C-kernel of MAGMA is much larger than that. Dozens of people from around the world have contributed code to MAGMA, at various levels, and there are unfortunately still many bugs. That said, the MAGMA developers are responsive and quickly fix the bugs and send out patches to users.

15.5 Implementations (probably goes on web page, not in book)

The following is MAGMA program that implements ECM.

```
// Returns either 2*P or GCD(N,x1-x2) =/= 1
function double(P,a,N)
   x,y,z := Explode(P);
   if z eq 0 then  // point at infinity
      return P;
   end if;
   g,_,y2inv := XGCD(N,Integers()!(2*y));
   if g ne 1 then
      return g;
   end if;
   xx := ((x^2-a)^2 - 8*x)*y2inv^2;
   yy := ((3*x^2 + a)*(x - xx) - 2*y^2)*y2inv;
   return [xx,yy,1];
end function;

// Returns P + Q or GCD(N,x1-x2) =/= 1
function add(P,Q,a,N)
```

```
   if P eq Q then
      return double(P,a,N);
   end if;
   x1,y1,z1 := Explode(P);
   x2,y2,z2 := Explode(Q);
   if z1 eq 0 then
      return Q;
   elif z2 eq 0 then
      return P;
   end if;
   if x1 eq x2 and y1 eq -y2 then
      return [0,1,0];
   end if;
   g,_,inv := XGCD(N,Integers()!(x1-x2));
   if g ne 1 then
      return g;
   end if;
   lambda := (y1-y2)*inv;
   nu := y1 - lambda*x1;
   x3 := lambda^2 -x1-x2;
   y3 :=  -lambda*x3-nu;
   return [x3,y3,1];
end function;

// Try to compute R=m*[0,1,1] on y^2=x^3+ax+1; returns
// either R or GCD(N,some denominator) =/= 1 if not possible.
function multiply(m,a,N)
   // Points are represented as triples [x,y,z] with z either 0 or 1.
   P := [IntegerRing(N)|0,1,1];
   R := [IntegerRing(N)|0,1,0];
   while m ne 0 do         // computes binary expansion of m.
      if IsOdd(m) then     // if binary digit of m is 1.
         R := add(R,P,a,N);
         if Type(R) eq RngIntElt then
            return R;
         end if;
      end if;
      m := Floor(m/2);
      P := double(P,a,N);
      if Type(P) eq RngIntElt then
         return P;
      end if;
   end while;
   return R;
end function;

intrinsic ECM1(N::RngIntElt, m::RngIntElt,
               a::RngIntElt) -> RngIntElt
{Try to find a B-power smooth factor of N using Lenstra's ECM
```

```
with given a and m=lcm(1,...,B). Returns N on failure.}
   printf "Trying a = %o: \t", a;
   if GCD(4*a^3 + 27, N) ne 1 then
      print "Split using discriminant.";
      return GCD(4*a^3 + 27, N);
   end if;
   R := multiply(m,a,N);
   if Type(R) eq RngIntElt then
      printf "Failed to compute mP. ";
      if R lt N then
         print "Split using denominator.";
         return R;
      end if;
         print "Denominator gives no factor.";
   end if;
   print "Computed mP (no factor found).";
   return N;
end intrinsic;


intrinsic ECM(N::RngIntElt, B::RngIntElt,
       maxtries::RngIntElt) -> RngIntElt
{Try to find a B-power smooth factor of N using Lenstra's ECM.
Returns N on failure. Stop after maxtries tries.}
   m := LCM([1..B]);
   for i in [1..maxtries] do
      a := Random(N);
      M := ECM1(N,m,a);
      if M ne N then
         return M;
      end if;
   end for;
   print "Max tries exceeded.  Trying changing B.";
   return N;
end intrinsic;
```

16

Solutions to Exercises

Chapter 3

1.1 First we show by induction on n that the binomial coefficients $\binom{n}{k}$, for $0 \leq k \leq n$, are integers. For the base case, note that $\binom{1}{0} = \binom{1}{1} = 1$. Next, suppose that $\binom{n}{k}$ is an integer for $0 \leq k \leq n$. Then $\binom{n+1}{0} = \binom{n+1}{n+1} = 1$, and for $0 \leq i \leq n-1$,

$$\begin{aligned}\binom{n}{i} + \binom{n}{i+1} &= \frac{n!}{i!(n-i-1)!}\left(\frac{1}{n-i} + \frac{1}{i+1}\right) \\ &= \frac{(n+1)!}{(i+1)!}(n-i)! = \binom{n+1}{i+1},\end{aligned}$$

which by the inductive hypothesis is the sum of two integers.

Since p is prime, it suffices to show that there is no factor of p in the denominator. By assumption, $r < p$ so $p \nmid r!$, and similarly $1 \leq r$ implies that $p - r \leq p - 1 < p$, so $(p-r)!$ also contains no factor of p.

1.2 The gcds are $5, 13, 3, 8$. E.g., $\gcd(15, 35) = \gcd(15, 5) = \gcd(5, 0) = \mathbf{5}$.

1.?? (a) The base case is trivial. If the statment is true for n, then

$$1 + \cdots + (n+1) = \frac{n(n+1)}{2} + (n+1) = \frac{(n+1)(n+2)}{2},$$

as desired.

(b) If n is even group the terms as $(1-2)+(3-4)+\cdots+(n-1-n) = -\frac{n}{2}$, and if n is odd we have $1 + (-2+3) + (-4+5) + \cdots + (-(n-1)+n) = \frac{n+1}{2}$, so the general formula is $(-1)^{n+1}\lceil \frac{n}{2} \rceil$.

1.5 The Euclidean algorithm gives $2261 = 1275 \cdot 1 + 986$, $1275 = 986 \cdot 1 + 289$, $986 = 289 \cdot 3 + 119$, $289 = 119 \cdot 2 + 51$, and $119 = 51 \cdot 2 + 17$, so

$$\begin{aligned} 17 &= 119 - 51 \cdot 2 = (986 - 289 \cdot 3) - (289 - 119 \cdot 2) \cdot 2 \\ &= 986 - (1276 - 986) \cdot 3 - (1275 - 986) \cdot 2 + (986 - 289 \cdot 3) \cdot 4 \\ &= 1275 \cdot (-5) + 986 \cdot 10 + 289 \cdot (-12) = 1275 \cdot (-17) + 986 \cdot 22 \\ &= \mathbf{22} \cdot 2261 - \mathbf{39} \cdot 1275. \end{aligned}$$

1.7 See Theorem 3.2.5 for the case $\deg(f) = 1$, and [Guy94, §A.1] when $\deg(f) = 2$. In general, if there is a prime p such that f induces the 0 function on $\mathbb{Z}/p\mathbb{Z}$, then $p \mid f(n)$ for all integers n, so f does not take on infinitely many values. Thus, for example, $f = x^2 + x + 2$ is irreducible, but takes on only finitely many prime values. If there is no such p and f is irreducible, then the author optimistically suspects that f takes on infinitely many prime values.

1.8 Easy.

1.9 First we note that all odd integers are congruent to ± 1 or 3 modulo 6, and if $x \equiv 3 \pmod 6$, then $3 \mid x$. Therefore all odd primes (except 3) are congruent to ± 1 modulo 6. Next we note that if $p, q \equiv 1 \pmod 6$, then $pq \equiv 1 \pmod 6$. Therefore if $n \equiv -1 \pmod 6$, then n must have a prime factor $p \equiv -1 \pmod 6$. Let $T = \{5, 11, \ldots\}$ be the set of all primes $p \equiv -1 \pmod 6$. If T is finite, it makes sense to consider the product Π_0 of the primes in T. Since $\Pi = 6\Pi_0 - 1 \equiv -1 \pmod 6$ there is a prime $q \in T$ such that $q \mid \Pi$. But for all $p \in T$, $\Pi \equiv -1 \pmod p$, a contradiction. Thus T is infinite.

1.10 (a) E.g., $\pi(2002) = 303$.

(b) For $x = 2001$, $x/\log(x) \approx 263$, so the values differ by about 40.

1.11 (a) $\{0, 1, 2, 3, 4, 5, 6\}$, (b) $\{1, 3, 5, 7, 9, 11, 13\}$, (c) $\{0, 2, 4, 6, 8, 10, 12\}$, and (d) $\{2, 3, 5, 7, 11, 13, 17\}$.

1.12 Here's the solution for divisibility by 11. *A positive number n is divisible by 11 if and only if the alternating sum of the digits of n is divisible by 11.* Proof. We use that $10 \equiv -1 \pmod{11}$ and facts about modular arithmetic. Write $n = \sum_{i=0}^{r} d_i 10^i$. Then $n \equiv \sum_{i=0}^{r} (-1)^i d_i \pmod{11}$, so $n \equiv 0 \pmod{11}$ if and only if the alternating sum of the digits $\sum (-1)^i d_i$ is congruent to 0 modulo 11.

1.13 71.

1.14 36.

1.15 $2^7 \cdot 3^2$.

1.17 302.

1.18 343.

1.19 61.

1.20 It suffices to prove that the congruence holds modulo each prime power $p_i^{r_i}$ appearing in the prime factorization $p_1^{r_1} \cdots p_m^{r_m}$ of n. If $m > 1$, then $p_i^{r_i} < n$, so $p_i^{r_i} \mid (n-1)!$, as required. It remains to consider the case when $m = 1$ and $n = p^r$ is a prime power. We have $2p \leq p^r - 1$ and $p^{r-1} \leq p^r - 1$, unless $p^r = 2^2$, which is not the case by assumption. Also, $2p \neq p^{r-1}$, unless $p = 2$ and $r = 3$, in which case one can check the statement we are proving directly. Thus

$$p^r \mid 2p \cdot p^{r-1} = 2p^r \mid (p^r - 1)!.$$

1.21 $\varphi(n)$ is odd only for $n = 1, 2$, as one sees by using the multiplicativity of φ and that if $p^r > 2$ is a prime power, then $\varphi(p^r) = p^r - p^{r-1} = (p-1)p^{r-1}$ is even.

1.24 For the first part, the answer is $n = 1, 2$. On a previous homework we proved that $n = 1, 2$ are the only n such that $\varphi(n)$ is odd. For the second part, the fact that φ is multiplicative means that if $\gcd(m, n) = 1$ then $\varphi(mn) = \varphi(m) \cdot \varphi(n)$. When $\gcd(m, n) \neq 1$ this implication can fail; for example, $2 = \varphi(2 \cdot 2) \neq \varphi(2) \cdot \varphi(2) = 1$.

Chapter 4

2.1 The secret code is $s = g^{mn} = 454^{1208} = 2156$ (all modulo 3793).

2.2 (a) 5384375093542

(b) 13, since the word with n Zs corresponds to $27^n - 1$ and 13 is the greatest integer n such that $27^n - 1 < 10^{20}$.

2.3 The numbers they chose are tiny, so we can write a program to very quickly compute n and m (simply try all posibilities ≤ 96). We find that $n = 70$ and $m = 31$, so $s = 5^{70 \cdot 31} = 44$.

2.5 (a) Using a computer, we find that $5352381469067 = 141307 \cdot 37877681$, so $\varphi(n) = (141307-1) \cdot (37877681-1) = 5352343450080$. The inverse of $e = 4240501142039$ modulo $\varphi(n)$ is $d = 5195621988839$.

(b) We have $3539014000459^d = 18464$, which encodes `WHY`.

2.6 Encrypting every letter individually is not a secure method. Rather, it is as secure as using a table of mappings of every letter to a number on both ends. This sort of encryption is very susceptible to word frequency examination.

2.7 (a) Let $m = ed - 1 = 38334587566167741692779486096$. We find that $a^{m/2} \equiv 1 \pmod{n}$ for several random choices of a. Replace m by $m/2 = 19167293783083870846389743048$. Again, we find that $a^{m/2} \equiv 1 \pmod{n}$ for several random a. Upon replacing m by $m/2 = 9583646891541935423194871524$, we find that $2^{m/2} \equiv 1433811615146880 \pmod{n}$. Unfortunately, $\gcd(2^{m/2} - 1, n) = 1$, so we try $3^{m/2}$, $5^{m/2}$, and finally find that $p = \gcd(7^{m/2} - 1) = 37865717$, which is a nontrivial factor of n. Setting $q = n/p = 37865693$, we have that $n = pq$.

(b) Letting $t = \lfloor\sqrt{n}\rfloor = 37865704$, we have $\sqrt{t^2 - n} \approx 8702.37$ and $\sqrt{(t+1)^2 - n} \approx 12.00$, so already we see that $p = t + 1 + 12 = 37865717$ and $q = t + 1 - 12 = 37865693$.

2.8 Since $2^n \equiv 6 \pmod{13}$, a table of powers of 2 modulo 13 quickly reveals that n must be 5 (we solve the discrete log problem easily in this case since 13 is so small). Likewise, since $g^m \equiv 11 \pmod{13}$ we see that $m = 7$. The secret key is $s = 7$ since $g^{nm} = 2^{35} \equiv 2^{11} \equiv 7 \pmod{13}$.

2.9 (a) We must compute $4^7 \pmod{77}$. Working modulo 77, we have that

$$4^7 = 64 \cdot 64 \cdot 4 = 13^2 \cdot 4 = 169 \cdot 4 = 15 \cdot 4 = 60,$$

so 4 encrypts as 60.

(b) First, $\varphi(n) = \varphi(77) = \varphi(7) \cdot \varphi(11) = 6 \cdot 10 = 60$. We then use the extended Euclidean algorithm to find an integer e such that $7e \equiv 1 \pmod{60}$. We find that $2 \cdot 60 - 17 \cdot 7 = 1$, so $e = -17$ is a solution.

Chapter 6

3.1 $\left(\frac{3}{97}\right) = (-1)^{48} \cdot \left(\frac{97}{3}\right) = \left(\frac{1}{3}\right) = \mathbf{1}$, $\left(\frac{5}{389}\right) = (-1)^2 \cdot \left(\frac{389}{5}\right) = \left(\frac{4}{5}\right) = \mathbf{1}$, $\left(\frac{2003}{11}\right) = \left(\frac{1}{11}\right) = \mathbf{1}$, and $\left(\frac{5!}{7}\right) = \left(\frac{120}{7}\right) = \left(\frac{1}{7}\right) = \mathbf{1}$.

3.2 By quadratic reciprocity $\left(\frac{3}{p}\right) = (-1)^{\frac{p-1}{2}} \cdot \left(\frac{p}{3}\right)$. Since $(-1)^{\frac{p-1}{2}}$ only depends on $p \pmod 4$ and $\left(\frac{p}{3}\right)$ only depends on $p \pmod 3$, their product depends only on $p \pmod{12}$, and the result follows by checking that the statement is true for each possibility of p modulo 12.

3.?? It is sufficient to give two distinct elements a, b in $(\mathbb{Z}/2^n\mathbb{Z})^\times$ of order 2, for if there was a primitive root g, then $g^{\phi(2^n)/2} = g^{2^{n-2}}$ cannot simultaneously be congruent to a and b modulo 2^n. Put $a = -1$; since $n > 2$, -1 has order 2 in $(\mathbb{Z}/2^n\mathbb{Z})^\times$. Set $b = 2^{n-1} - 1 \in (\mathbb{Z}/2^n\mathbb{Z})^\times$; then $b^2 = (2^{n-1} - 1)^2 = 2^{2n-2} - 2 \cdot 2^{n-1} + 1 = 1$. Now $b \neq a$, since $b - a = 2^{n-1} < 2^n$, and $b \neq 1$, since $n > 2$. Therefore a and b are distinct elements of order 2 in $(\mathbb{Z}/2^n\mathbb{Z})^\times$, a contradiction.

3.?? We will construct an element g of $(\mathbb{Z}/p^2\mathbb{Z})^\times$ with order $\varphi(p^2) = p(p-1)$. Let g_0 be a primitive root modulo p, and let $g = g_0 + pt$ for some t to be determined. By the binomial theorem

$$g^{p-1} \equiv g_0^{p-1} + (p-1)pg_0^{p-2}t \equiv (1 + kp) + p(p-1)g_0^{p-2}t \pmod{p^2},$$

for some k, since $g_0^{p-1} \equiv 1 \pmod p$. Because $p - 1$ and g_0^{p-2} are both coprime to p, we can choose t such that $n = k + (p-1)g_0^{p-2}t$ is nonzero modulo p. Then $g^{p-1} \equiv 1 + np \pmod{p^2}$ and $p \nmid n$, so the order of g in $(\mathbb{Z}/p^2\mathbb{Z})^\times$ does not divide $p - 1$. But it divides $p(p-1)$, and p is prime, so the order of g is $p(p-1)$, and g is a primitive root modulo p^2.

3.3 Let g be a primitive root modulo p. Since $p \equiv 1 \pmod 3$, $c = g^{(p-1)/3}$ has order 3. Therefore c is a solution to

$$x^3 - 1 = (x-1)(x^2+x+1) = 0,$$

where all arithmetic is taking place in the integral domain $\mathbb{Z}/p\mathbb{Z}$. Since $c \neq 1$, we have $c^2 + c + 1 = 0$. Thus $4c^2 + 4c + 4 = (2c+1)^2 + 3 = 0$, so $(2c+1)^2 = -3$, hence $\left(\frac{-3}{p}\right) = 1$.

3.4 The solution is similar to the solution of the Exercise 3.4. Let c in $(\mathbb{Z}/p\mathbb{Z})^\times$ be an element of order 5. Then $c^5 - 1 = (c-1)(c^4 + c^3 + c^2 + c + 1) = 0$ and $c \neq 1$ implies that $c^4 + c^3 + c^2 + c + 1 = 0$. Now

$$(c+c^4)^2 + (c+c^4) - 1 = c^4 + c^3 + c^2 + c + 1 = 0,$$

so $(2(c+c^4)+1)^2 = 4((c+c^4)^2 + (c+c^4) - 1) + 5 = 5$.

3.5 *All odd primes.* Let p be an odd prime and g a primitive root modulo p. Rewrite the sum as:

$$\begin{aligned}\sum_{a=1}^{p-1}\left(\frac{a}{p}\right) &= \sum_{i=1}^{p-1}\left(\frac{g^i}{p}\right) = \sum_{j=1}^{\frac{p-1}{2}}\left(\frac{g^{2j}}{p}\right) + \sum_{j=1}^{\frac{p-1}{2}}\left(\frac{g^{2j+1}}{p}\right)\\ &= \sum_{j=1}^{\frac{p-1}{2}}\left(\frac{g^{2j}}{p}\right) + \left(\frac{g}{p}\right)\sum_{j=1}^{\frac{p-1}{2}}\left(\frac{g^{2j}}{p}\right) = \frac{p-1}{2}\left(1 + \left(\frac{g}{p}\right)\right).\end{aligned}$$

Now $\left(\frac{g}{p}\right) = -1$, for if $\left(\frac{g}{p}\right) = 1$ then $g^{\frac{p-1}{2}} = 1$, and g would not be primitive. Therefore $\sum_{a=1}^{p-1}\left(\frac{a}{p}\right) = 0$.

3.?? A good guess is to be $C \approx 0.37395$. Using a computer, we can write a program to check the first n primes to see if 2 is a primitive root, and divide this total by n to see the behavior of the ratio. Artin conjectured that

$$C = \prod_{n=1}^{\infty}\left(1 - \frac{1}{p_n(p_n-1)}\right),$$

where p_n is the nth prime.

3.6 First we use the law of quadratic reciprocity to decide whether or not there is a solution. We have

$$\left(\frac{5}{2^{13}-1}\right) = (-1)^{2\cdot(2^{13}-2)/2}\left(\frac{2^{13}-1}{5}\right) = \left(\frac{1}{5}\right) = 1,$$

so the equation $x^2 \equiv 5 \pmod{2^{13}-1}$ has at least one solution a. Since the polynomial $x^2 - 5$ has degree two and $2^{13}-1$ is prime, there are at most 2 solutions. Since $-a$ is also a solution and $a \neq 0$, there are *exactly two solutions.*

3.7 Fermat's Little Theorem implies that $4^{48} = 2^{96} \equiv 1 \pmod{97}$.

Chapter 7

4.2 We establish both identities by induction.

- Claim: $[a_n, a_{n-1}, \ldots, a_1, a_0] = \frac{p_n}{p_{n-1}}$.

  Proof. Since $p_{-1} = 1$ and $p_0 = a_0$, $[a_0] = \frac{p_0}{p_{-1}}$, which establishes the base case. Now suppose that $[a_{n-1}, \ldots, a_0] = \frac{p_{n-1}}{p_{n-2}}$. Then

$$\begin{aligned}[a_n, a_{n-1}, \ldots, a_0] &= a_n + \frac{1}{[a_{n-1}, \ldots, a_0]} \\ &= a_n + \frac{1}{\frac{p_{n-1}}{p_{n-2}}} = a_n + \frac{p_{n-2}}{p_{n-1}} = \frac{a_n p_{n-1} + p_{n-2}}{p_{n-1}}.\end{aligned}$$

  The numerator is p_n, by definition. □

- Claim: $[a_n, a_{n-1}, \ldots, a_1] = \frac{q_n}{q_{n-1}}$.

  Proof. Since $q_0 = 1$ and $q_1 = a_1$, $[a_1] = \frac{q_1}{q_0}$, which establishes the base case. Now suppose that $[a_{n-1}, \ldots, a_1] = \frac{q_{n-1}}{q_{n-2}}$. Then

$$\begin{aligned}[a_n, a_{n-1}, \ldots, a_1] &= a_n + \frac{1}{[a_{n-1}, \ldots, a_1]} \\ &= a_n + \frac{1}{\frac{q_{n-1}}{q_{n-2}}} = a_n + \frac{q_{n-2}}{q_{n-1}} = \frac{a_n q_{n-1} + q_{n-2}}{q_{n-1}}.\end{aligned}$$

  The numerator is q_n, by definition. □

4.?? If we compute `ellj(t)` in PARI, where $t = -0.5 + 0.3281996289i$ the result is `61.7142856...-6.2E-26I`. The imaginary part is approximately 0, so we guess the rational number that gives the real part. The command `contfrac(61.7142856)` gives `[61,1,2,2,178571]`, which suggests that a good guess for our rational number is the rational number $\frac{432}{7}$ represented by the continued fraction $[61, 1, 2, 2]$.

4.3 (a) Let $\alpha = [\overline{2,3}]$. Then $\alpha = 2 + \frac{1}{3+\frac{1}{\alpha}}$, so $3\alpha^2 - 6\alpha - 2 = 0$. Solving for α yields $1 + \frac{\sqrt{15}}{3}$.

(b) First we compute $\alpha = [\overline{1,2,1}]$. This gives $\alpha = 1 + \left(2 + \frac{1}{(1+\alpha^{-1})}\right)^{-1}$. Solving for α yields $3\alpha^2 - 2\alpha - 3 = 0$, so $\alpha = \frac{1+\sqrt{10}}{3}$. Now $[2, \overline{1,2,1}] = 2 + \frac{1}{[\overline{1,2,1}]}$, so our final answer is $\frac{5+\sqrt{10}}{3}$.

(c) This is $[\overline{1,2,3}]^{-1}$. As above, if $\alpha = [\overline{1,2,3}]$ then $\alpha = 1 + \left(2 + \frac{1}{3+\frac{1}{\alpha}}\right)^{-1}$. This simplifies to $\alpha = 7\alpha^2 + 8\alpha - 3 = 0$, so $\alpha = \frac{4+\sqrt{37}}{7}$. Therefore our desired answer is $\frac{-4+\sqrt{37}}{3}$.

4.4 We use the command `contfrac` in PARI to find the continued fraction then prove that the answer is correct.

(a) We claim that $\sqrt{5} = [2, \overline{4}]$. Let $\alpha = [\overline{4}]$; then $\alpha = 4 + \frac{1}{\alpha}$, so $\alpha = 2 + \sqrt{5}$. Now $[2, \overline{4}] = 2 + \frac{1}{2+\sqrt{5}} = \sqrt{5}$, as desired.

(b) We claim that $\frac{1+\sqrt{13}}{2} = [2, \overline{3}]$. Let $\alpha = [\overline{3}]$; then $\alpha = 3 + \frac{1}{\alpha}$, so $\alpha^2 - 3\alpha - 1 = 0$. This gives $\alpha = \frac{3+\sqrt{13}}{2}$. Then $[2, \overline{3}] = 2 + \frac{2}{3+\sqrt{13}} = \frac{1+\sqrt{13}}{2}$.

(c) We claim that $\frac{5+\sqrt{37}}{4} = [\overline{2,1,3}]$. Let $\alpha = [\overline{2,1,3}]$; then $\alpha = 2 + \left(1 + (3 + \frac{1}{\alpha})^{-1}\right)^{-1}$, so $4\alpha^2 - 10\alpha - 3 = 0$. Solving for α gives $\frac{5+\sqrt{37}}{4}$, as desired.

4.5 (a) First we compute $[\overline{2n}]$. Let $\alpha = [\overline{2n}]$; then $\alpha = 2n + \frac{1}{\alpha}$. This gives $\alpha^2 - 2n\alpha - 1 = 0$, so $\alpha = n + \sqrt{n^2+1}$. Now $[n, \overline{2n}] = n + \frac{1}{\alpha} = n - (n - \sqrt{n^2+1}) = \sqrt{n^2+1}$, as desired.

(b) Using the previous part, we know that $\sqrt{5} = [2, \overline{4}]$. We can try successive convergents until two agree up to four decimal places; once such convergent is $682/305$.

4.6 In PARI, use `convergents(contfrac(Pi))` to obtain the convergents of the continued fraction of π. We can now test convergents for property described in the problem, noting that smaller denominators are more likely to work. One convergent that satisfies the property is $3/1$, since $\pi - 3 < .15$ and $\frac{1}{\sqrt{5}} > .44$. The next is $22/7$, since $22/7 - \pi < .002$ and $\frac{1}{49\sqrt{5}} > .009$. A third is $355/113$, since $355/113 - \pi < .0000003$ while $\frac{1}{113^2\sqrt{5}} > .00003$.

4.7 In PARI, the command `contfrac(exp(2))` gives

```
[7,  2,1,1,3,18,   5,1,1,6,30,   8,1,1,9,42,  11,1,1,12,54,
     14,1,1,15,77, 17,1,1,18,78, 20,1,1,21,90,...]
```

This suggests that after the initial 7 terms, the ith group of 5 numbers (grouping after the initial 7) is of the form

$$3(i-1) + 2, 1, 1, 3i, 18 + 12(i-1).$$

4.8 (a) We are looking for natural number solutions to $n + 1 = x^2$, $\frac{n}{2} + 1 = y^2$. Isolating n yields $n = x^2 - 1 = 2(y^2 - 1)$, implying that $x^2 - 2y^2 = -1$. There are infinitely many solutions (x, y) to this equation if the period of the continued fraction of $\sqrt{2}$ has odd order. Indeed, $\sqrt{2} = [1, \overline{2}]$. For each (x, y) we can take $2(y^2 - 1)$ to find a unique (even) n; thus there are infinitely many n satisfying the desired property.

(b) We compute list of convergents for the continued fraction of $\sqrt{2}$. The odd terms are of interest, and since we want to find $n > 389$, we want the denominator of the convergent to be at least 14. The first two such convergents are $\frac{41}{29}$ and $\frac{239}{169}$, which yield $n = 2(29^2 - 1) = 1680$ and $n = 2(169^2 - 1) = 57120$.

4.9 If x and y are consecutive integers, then we have $z^2 = x^2 + (x+1)^2 = 2x^2 + 2x + 1$. Multiplying by 2, we have $2z^2 = 4x^2 + 4x + 2 = (2x+1)^2 + 1$. Putting $u = 2x + 1$, we have $u^2 - 2z^2 = -1$. From the previous problem we know there are infinitely many solutions to this equation for u and z. Each solution (u, z) gives a unique (x, z), and each of these solutions (x, z) leads to a primitive Pythagorean triple (x and $x+1$ are coprime and if either shares a nontrivial divisor with z then the third must also share this divisor).

Chapter 9

5.1 Since -389 is negative, it is not the sum of two squares. Since $3 \,\|\, 12345$, we see that 12345 is not the sum of two squares. Since $7 \,\|\, 91210$, it follows that 91210 is not the sum of two squares. The fourth number is $729 = 27^2$, which is a perfect square. Since $7 \,\|\, 1729$, we see that 1729 is not the sum of two squares. Finally, $151 \,\|\, 68252$ and $151 \equiv 3 \pmod 4$, so 68252 is not the sum of two squares.

5.2 (a) Given an input n, the following PARI program breaks n up into two parts and looks for a sum of two squares representation.

```
{squares(n) =
   local(y);
   for(x=1,floor(sqrt(n)),
      y=sqrt(n-x^2);
      if(y-floor(y)==0,
         return([x,floor(y)])
      )
   );
   return(0)
};

{f(n) =
   for(x=1,n,
      a=squares(x);
      b=squares(n-x);
      if(a!=0 && b!=0,
         return([a[1],a[2],b[1],b[2]])
      )
   )
}
```

(b) $2001 = 1^2 + 8^2 + 44^2$

5.3 Answer: 625. There are two Pythagorean triples with 25 as the hypotenuse: $(7, 24, 25)$ and $(15, 20, 25)$. This gives two representations of 625 as the sum of two squares. Of course, 25^2 is a third.

5.4 The forward direction is trivial. For the opposite direction, suppose that n is the sum of two rational squares: $n = (\frac{a}{b})^2 + (\frac{c}{d})^2$, but that n

is not the sum of two integer squares. Then there exists a prime p such that $p^r \parallel n$, where $p \equiv 3 \pmod 4$ and r is odd. Now, $nb^2d^2 = (ad)^2 + (bc)^2$, so nb^2d^2 is the sum of two integer squares. However, all the prime factors of b^2d^2 have even exponent, so $p^s \parallel nb^2d^2$, where s is odd, a contradiction.

5.5 Suppose $p = x^2 + 2y^2$, where p is an odd prime and x and y are integers. Then $x^2 + 2y^2 \equiv 0 \pmod p$, so $\left(\frac{x}{y}\right)^2 \equiv -2 \pmod p$ (since $\mathbb{Z}/p\mathbb{Z}$ is a field). From XXX, $\left(\frac{2}{p}\right) = 1$ if and only if $p \equiv \pm 1 \pmod 8$. Also, $\left(\frac{-1}{p}\right) = 1$ if and only if $p \equiv 1 \pmod 4$ (i.e., $p \equiv 1, -3 \pmod 8$). Since $\left(\frac{-2}{p}\right) = \left(\frac{-1}{p}\right) \cdot \left(\frac{2}{p}\right)$, we have $\left(\frac{-2}{p}\right) = 1$ if and only if $p \equiv 1, 3 \pmod 8$.

Conversely, suppose that $p \equiv 1, 3 \pmod 8$ is prime. Let r be such that $r^2 \equiv -2 \pmod p$. Taking $n = \lfloor \sqrt{p} \rfloor$ and applying Lemma 1.3 from Lecture 21 (XXX), there exist integers a, b with $0 < b < \sqrt{p}$ such that

$$\left| -\frac{r}{p} - \frac{a}{b} \right| \le \frac{1}{b(n+1)} < \frac{1}{b\sqrt{p}}.$$

Let $c = rb + pa$; then $|c| < \frac{pb}{b\sqrt{p}} = \sqrt{p}$, so $2b^2 + c^2 < 3p$. Since $c \equiv rb \pmod p$, we also have that $2b^2 + c^2 \equiv b^2(2 + r^2) \equiv 0 \pmod p$. Therefore $2b^2 + c^2 = p$ or $2p$. If $2b^2 + c^2 = p$ we are done. If $2b^2 + c^2 = 2p$ then c must be even (else $2b^2 + c^2$ is odd). Putting $c = 2d$, we have

$$2p = 2b^2 + c^2 = 2b^2 + 4d^2,$$

so $p = b^2 + 2d^2$, as desired.

5.7 Let T_m be the mth triangular number. It is easy to see by induction that $T_m = m(m+1)/2$. Then

$$\begin{aligned}
8T_m^2 + 0 &= (2T_m)^2 + (2T_m)^2, \\
8T_m^2 + 1 &= 2m^2(m+1)^2 + 1 = [(m-1)(m+1)]^2 + [m(m+2)]^2, \\
8T_m^2 + 2 &= 2m^2(m+1)^2 + 2 = [m(m+1) - 1]^2 + [m(m+1) + 1]^2.
\end{aligned}$$

5.8 Of any four consecutive integers, there is one n such that $n \equiv -1 \pmod 4$. Since all odd prime factors are congruent to $\pm 1 \pmod 4$, n must have some prime factor $p \equiv -1 \pmod 4$ with odd exponent, so n is not representable as the sum of two squares.

5.9 We first solve

$$\begin{aligned}
13x^2 + 36xy + 25y^2 &= (ax + by)^2 + (cx + dy)^2 \\
&= (a^2 + c^2)x^2 + 2(ab + cd)xy + (b^2 + d^2)y^2,
\end{aligned}$$

then check that $ad - bc = 1$. By inspection, we try $a = 3, c = 2$. Then $b^2 + d^2 = 25$ and $2b + 3d = 18$, which again by inspection is satisfied by $b = 4, d = 3$. Since $ad - bc = 1$, the form is equivalent to $x^2 + y^2$.

As above, we solve

$$\begin{aligned} 58x^2 + 82xy + 29y^2 &= (ax + by)^2 + (cx + dy)^2 \\ &= (a^2 + c^2)x^2 + 2(ab + cd)xy + (b^2 + d^2)y^2, \end{aligned}$$

then check that $ad - bc = 1$. By inspection, we try $a = 3, c = 7$. Then $b^2 + d^2 = 29$ and $7b + 3d = 41$, which again by inspection is satisfied by $b = 2, d = 5$. Since $ad - bc = 1$, the form is equivalent to $x^2 + y^2$.

We know that $x = 17, y = 10$ satisfies $x^2 + y^2 = 389$. To find x, y such that $13x^2 + 36xy + 25y^2$, We use the transformation above and solve for

$$17 = 3x + 4y, \ 10 = 2x + 3y.$$

The solution to this system is $x = 11$, $y = -4$. Indeed,

$$13 \cdot 11^2 - 36 \cdot 11 \cdot 4 + 25 \cdot 4^2 = 389.$$

5.10 The discriminants are equal: $-24 = 162^2 - 4 \cdot 199 \cdot 33 = 96^2 - 4 \cdot 35 \cdot 66$. However, the forms are not equivalent. To see this, we first show that $35x^2 - 96xy + 66y^2$ is equivalent to $2x^2 + 3y^2$. As above, this means solving

$$\begin{aligned} 35x^2 - 96xy + 66y^2 &= 2(ax + by)^2 + 3(cx + dy)^2 \\ &= (2a^2 + 3c^2)x^2 + 2(2ab + 3cd)xy + (2b^2 + 3d^2)y^2 \end{aligned}$$

in integers a, b, c, d such that $ad - bc = 1$. By inspection we see that $a = 2$, $b = -3$, $c = 3$, $d = -4$ is a solution, so the forms are equivalent. Now we show the first form is not equivalent to $2x^2 + 3y^2$. If we try to solve for as above, we encounter the equation $33 = 2b^2 + 3d^2$, which has no solutions over the integers (we can just check $0 \le b \le 5$).

5.11

- $D = -155$: The group C_D has four elements, represented by $[[1, 1, 39], [3, -1, 13], [3, 1, 13], [5, 5, 9]]$. By the structure theorem, $\mathcal{C}_D$ is isomorphic to either $C_2 \times C_2$ or C_4. It is easy to verify that $[1, 1, 39]$ is the identity. From this we find that $[3, -1, 13]$ has order 4, so $\mathcal{C}_D \approx C_4$.
- $D = -231$: There are 12 elements: $[1, 1, 58]$, $[2, -1, 29]$, $[2, 1, 29]$, $[3, 3, 20]$, $[4, -3, 15]$, $[4, 3, 15]$, $[5, -3, 12]$, $[5, 3, 12]$, $[6, -3, 10]$, $[6, 3, 10]$, $[7, 7, 10]$, $[8, 5, 8]$, so $\mathcal{C}_D \approx C_{12}$ or $C_2 \times C_6$. The identity is $[1, 1, 58]$ and both $[2, -1, 29]$ and $[2, 1, 29]$ have order 6, which is impossible in C_{12}, so $\mathcal{C}_D \approx C_2 \times C_6$.
- $D = -660$: There are eight elements: $[1, 0, 165]$, $[10, 10, 19]$, $[11, 0, 15]$, $[13, 4, 13]$, $[2, 2, 83]$, $[3, 0, 55]$, $[5, 0, 33]$, $[6, 6, 29]$. The first element is the identity, and all others have order 2, so $\mathcal{C}_{-660} \approx C_2 \times C_2 \times C_2$.
- $D = -12104$: There are forty-eight elements (and these can take a while to compute!): By the structure theorem,

$$\mathcal{C}_D \approx C_{48}, C_4 \times C_{12}, \text{ or } C_2 \times C_{24}.$$

The identity element is $[1, 0, 3026]$, and using it we find two elements of order four: $[45, -26, 71]$ and $[50, -36, 67]$, eliminating everything but $C_4 \times C_{12}$.

- $D = -10015$: There are fifty-four elements (which take a while to compute!). Therefore $\mathcal{C}_D \approx C_3 \times C_{18}$ or C_{54}. The identity is $[1, 1, 2504]$, and from this we find that $[10, -5, 251]$ and $[10, 5, 251]$ have order 9. Thus $\mathcal{C}_D$ cannot be C_{54}, so $\mathcal{C}_D \approx C_3 \times C_{18}$.

Chapter 10

6.6

6.7 Differentiating implicitly, the slope of the tangent at (x, y) is $\frac{3x^2}{2y}$. At $(3, 5)$, the slope is $\frac{27}{10}$, and the tangent line has equation $y = \frac{27x-31}{10}$. Substituting into the relation $y^2 - x^3 = -2$, we have $\left(\frac{27x-31}{10}\right)^2 - x^3 = -2$, which simplifies to the polynomial

$$100x^3 - 729x^2 + 1674x - 1161 = 0.$$

This polynomial has a double root at $x = 3$, so it factors as $(100x - 129)(x-3)^2$, giving a rational root with $x = 129/1000$. Therefore $(129/100, 383/1000)$ is a rational solution to the original equation.

6.9 The transformation

$$x \mapsto X - \frac{1}{12}, \qquad y \mapsto -\frac{1}{2}X + Y - \frac{11}{24}.$$

transforms $y^2 + xy + y = x^3$ into

$$Y^2 = X^3 + \frac{23}{48}X + \frac{181}{864}.$$

6.11 (a) The elements are $\{\mathcal{O}, (0, 1), (0, 4), (2, 1), (2, 4), (3, 1), (3, 4), (4, 2), (4, 3)\}$.

(b) The element $(0, 1)$ has order 9; thus, $E(K) \cong C_9$.

6.12 (a) In general there is no identity element. Suppose to the contrary that $\mathcal{O}$ is the identity, and suppose that P, Q, and $\mathcal{O}$ are distinct and colinear. Then $P \boxplus \mathcal{O} = Q \neq P$, so $\mathcal{O}$ is not the identity. The inverse axiom sort of works, in that if $\mathcal{O}$ were the identity and $P \in E(\mathbb{R})$, then there is Q such that $P + Q = \mathcal{O}$. Associativity fails; for example, consider $y^2 = x^3 + 1$. We have $((-1, 0) \boxplus (0, 1)) \boxplus (0, -1) = (2, 3)$, yet $(-1, 0) \boxplus ((0, 1) \boxplus (0, -1)) = (-1, 0)$.

(b) Note that $P \boxplus Q = -(P + Q)$, where $+$ is the usual group law. Thus if $E(\mathbb{Q}) \in \{\{0\}, \mathbb{Z}/2\mathbb{Z}, (\mathbb{Z}/2\mathbb{Z})^2\}$, then $(E(\mathbb{Q}), \boxplus)$ is a group. Otherwise it isn't, because there is no identity, as mentioned above.

6.13 (a) Substituting the equations for x and y into $y^2 = f(x)$ and isolating u^2 yields:

$$u^2 = g'(\alpha)(t-\alpha) + \frac{1}{2}g''(\alpha)(t-\alpha)^2 + \frac{1}{6}g'''(t-\alpha)^3 + \frac{1}{24}g''''(t-\alpha)^4$$

Since α is a root of $g(t)$ and $g(t)$ is a quartic polynomial, the right-hand side is precisely the same as the (entire) Taylor series of g around α. Thus, the right-hand side is equal to $g(t)$ itself.

(b) g has four distinct roots; call them x_1, x_2, x_3, x_4. Without loss of generality, let $x_1 = \alpha$; then $t = \frac{\beta}{x} + x_1$. Thus,

$$g(t) = \left(\frac{\beta}{x} + x_1 - x_1\right)\left(\frac{\beta}{x} + x_1 - x_2\right)\left(\frac{\beta}{x} + x_1 - x_3\right)\left(\frac{\beta}{x} + x_1 - x_4\right)$$

Multiplying by x^4 transforms the left-hand side into $f(x)$ and the right-hand side into a polynomial that clearly has distinct roots. This directly implies $u^2 = g(t)$ is an elliptic curve.

6.14 (a) Symmetry considerations ensure that the arc length of C is four times the arc length of $y = \beta\sqrt{1 - \frac{x^2}{a^2}}$ from $x = 0$ to $x = \alpha$. The arc length formula is:

$$\int \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx$$

so the requisite substitution gives:

$$S = 4\int_0^{\alpha} \sqrt{1 + \frac{\beta^2 x^2}{\alpha^2(\alpha^2 - x^2)}} dx$$

Substituting $\sin\theta = \frac{x}{\alpha}$ yields:

$$S = 4\alpha \int_0^{\frac{\pi}{2}} \sqrt{1 - \frac{\alpha^2 - \beta^2}{\alpha^2} \sin^2\theta d\theta}$$

Letting $k = \frac{\sqrt{\alpha^2 - \beta^2}}{\alpha^2}$ yields the desired result.

(b) If $\alpha = \beta$, the formula yields the correct answer of $S = 2\pi\alpha$.

(c) The result arises from a substitution of $t = \frac{x}{\alpha}$ into the arc length integral in part a.

(d) E is not a circle, so $\alpha \neq \beta$, and $k \neq 0$. Then the quartic polynomial $g(t) = (1 - t^2)(1 - k^2 t^2)$ has distinct roots $x = \pm 1, \pm\frac{1}{k}$. By the previous exercise, there exists a cubic polynomial $f(x)$ such that $y^2 = f(x)$ defines the elliptic curve associated with $g(t)$.

6.15

6.16

6.?? After transforming the curve into the form $Y^2 = X^3 + aX + b$ as per exercise 6.9, the formula for the discriminant gives:

$$\Delta = 1009310191436363411578571211896385085440000000 = 2^{18}{\cdot}3^{18}{\cdot}5^{6}{\cdot}13^{6}{\cdot}23^{6}{\cdot}61^{2}{\cdot}67^{2}{\cdot}73^{2}$$

In theory, using Lutz-Nagell in this case would entail checking all integral values of Y such that $Y^2 | \Delta$ to see if there was an integer X such that (X, Y) was on the curve, and if there was, to see if it had a finite order. This would yield the elements of the torsion subgroup, and Mazur's theorem would allow us to deduce its structure. However, in this particular case, the number of values of Y that would have to be checked is prohibitively large.

6.17

6.??

6.18 (a) A finite set of points on E can only generate a countable number of other points on E. $E(\mathbb{R})$ is uncountable; thus, it cannot be a finitely generated abelian group.

(b) $E(k)$ has only a finite number of elements; thus, it is finitely generated.

References

[ACD$^+$99] K. Aardal, S. Cavallar, B. Dodson, A. Lenstra, W. Lioen, P. L. Montgomery, B. Murphy, J. Gilchrist, G. Guillerm, P. Leyland, J. Marchand, F. Morain, A. Muffett, C.&C. Putnam, and P. Zimmermann, *Factorization of a 512-bit RSA key using the Number Field Sieve*, `http://www.loria.fr/~zimmerma/records/RSA155` (1999).

[Ahl78] L. V. Ahlfors, *Complex analysis*, third ed., McGraw-Hill Book Co., New York, 1978, An introduction to the theory of analytic functions of one complex variable, International Series in Pure and Applied Mathematics. MR 80c:30001

[AKS02] M. Agrawal, N. Kayal, and N. Saxena, *Primes is in p*, `http://www.cse.iitk.ac.in/users/manindra/` (2002).

[BCP97] W. Bosma, J. Cannon, and C. Playoust, *The Magma algebra system. I. The user language*, J. Symbolic Comput. **24** (1997), no. 3-4, 235–265, Computational algebra and number theory (London, 1993). MR 1 484 478

[Ber] D. J. Bernstein, *An Exposition of the Agrawal-Kayal-Saxena Primality-Proving Theorem*, `http://cr.yp.to/papers/aks.ps`.

[Bur89] David M. Burton, *Elementary number theory*, second ed., W. C. Brown Publishers, Dubuque, IA, 1989. MR 90e:11001

[Cal] C. Caldwell, *The largest known primes*, `http://www.utm.edu/research/primes/largest.html`.

[Cas62] J. W. S. Cassels, *Arithmetic on Curves of Genus* 1*. IV. Proof of the Hauptvermutung*, J. Reine Angew. Math. **211** (1962), 95–112.

[Coh93] H. Cohen, *A course in computational algebraic number theory*, Graduate Texts in Mathematics, vol. 138, Springer-Verlag, Berlin, 1993. MR 94i:11105

[CP01] R. Crandall and C. Pomerance, *Prime numbers*, Springer-Verlag, New York, 2001, A computational perspective. MR 2002a:11007

[Dav99] H. Davenport, *The higher arithmetic*, seventh ed., Cambridge University Press, Cambridge, 1999, An introduction to the theory of numbers, Chapter VIII by J. H. Davenport. MR 2000k:11002

[DI95] F. Diamond and J. Im, *Modular forms and modular curves*, Seminar on Fermat's Last Theorem, Providence, RI, 1995, pp. 39–133.

[FT93] A. Fröhlich and M. J. Taylor, *Algebraic number theory*, Cambridge University Press, Cambridge, 1993. MR 94d:11078

[GS02] X. Gourdon and P. Sebah, *The pi(x) project*, `http://numbers.computation.free.fr/Constants/Primes/Pix/pixproject.html` (2002).

[Guy94] R.K. Guy, *Unsolved problems in number theory*, second ed., Springer-Verlag, New York, 1994, Unsolved Problems in Intuitive Mathematics, I. MR 96e:11002

[Hal60] P.R. Halmos, *Naive set theory*, The University Series in Undergraduate Mathematics, D. Van Nostrand Co., Princeton, N.J.-Toronto-London-New York, 1960. MR 22 #5575

[Hoo67] Christopher Hooley, *On Artin's conjecture*, J. Reine Angew. Math. **225** (1967), 209–220. MR 34 #7445

[HW79] G.H. Hardy and E.M. Wright, *An introduction to the theory of numbers*, fifth ed., The Clarendon Press Oxford University Press, New York, 1979. MR 81i:10002

[IBM01] IBM, *IBM's Test-Tube Quantum Computer Makes History*, `http://www.research.ibm.com/resources/news/20011219_quantum.shtml` (2001).

[IR90] K. Ireland and M. Rosen, *A classical introduction to modern number theory*, second ed., Springer-Verlag, New York, 1990. MR 92e:11001

[Khi63] A. Ya. Khintchine, *Continued fractions*, Translated by Peter Wynn, P. Noordhoff Ltd., Groningen, 1963. MR 28 #5038

[Leh14] D.N. Lehmer, *List of primes numbers from* 1 *to* 10,006,721, Carnegie Institution Washington, D.C. (1914).

[Lem] F. Lemmermeyer, *Proofs of the Quadratic Reciprocity Law*, `http://www.rzuser.uni-heidelberg.de/~hb3/rchrono.html`.

[Len87] H.W. Lenstra, Jr., *Factoring integers with elliptic curves*, Ann. of Math. (2) **126** (1987), no. 3, 649–673. MR 89g:11125

[Len02] ———, *Solving the Pell equation*, Notices Amer. Math. Soc. **49** (2002), no. 2, 182–192. MR 2002i:11028

[LT72] S. Lang and H. Trotter, *Continued fractions for some algebraic numbers*, J. Reine Angew. Math. **255** (1972), 112–134; addendum, ibid. **267** (1974), 219–220; MR **50** #2086. MR 46 #5258

[LT74] ———, *Addendum to: "Continued fractions for some algebraic numbers" (J. Reine Angew. Math.* **255** *(1972), 112–134)*, J. Reine Angew. Math. **267** (1974), 219–220. MR 50 #2086

[Mat] Matsucom, *The on*hand *PC Watch*, `http://www.pconhand.com/`.

[Mor93] Pieter Moree, *A note on Artin's conjecture*, Simon Stevin **67** (1993), no. 3-4, 255–257. MR 95e:11106

[MW00] R. Martin and McMillen W., *An Elliptic Curve over* **q** *with Rank at least 24*, `http://listserv.nodak.edu/scripts/wa.exe?A2=ind0005&L=nmbrthry&P=R182` (2000).

[Per50] O. Perron, *Die Lehre von den Kettenbrüchen*, Chelsea Publishing Co., New York, N. Y., 1950, 2d ed. MR 12,254b

[Rib90] K. A. Ribet, *On modular representations of* $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ *arising from modular forms*, Invent. Math. **100** (1990), no. 2, 431–476.

[RSA] RSA, *The New RSA Factoring Challenge*, `http://www.rsasecurity.com/rsalabs/challenges/factoring`.

[RSA78] R. L. Rivest, A. Shamir, and L. Adleman, *A method for obtaining digital signatures and public-key cryptosystems*, Comm. ACM **21** (1978), no. 2, 120–126. MR 83m:94003

[Ser73] J-P. Serre, *A Course in Arithmetic*, Springer-Verlag, New York, 1973, Translated from the French, Graduate Texts in Mathematics, No. 7.

[Sho97] P. W. Shor, *Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer*, SIAM J. Comput. **26** (1997), no. 5, 1484–1509. MR 98i:11108

[Sil86] J. H. Silverman, *The arithmetic of elliptic curves*, Graduate Texts in Mathematics, vol. 106, Springer-Verlag, New York, 1986. MR 87g:11070

[Sin97] S. Singh, *The Proof*, `http://www.pbs.org/wgbh/nova/transcripts/2414proof.html` (1997).

[Sin99] ———, *The Code Book: The Science of Secrecy from Ancient Egypt to Quantum Cryptography*, Doubleday, 1999.

[ST92] J. H. Silverman and J. Tate, *Rational points on elliptic curves*, Undergraduate Texts in Mathematics, Springer-Verlag, New York, 1992. MR 93g:11003

[Sta78] Harold M. Stark, *An introduction to number theory*, MIT Press, Cambridge, Mass., 1978. MR 80a:10001

[Wal48] H. S. Wall, *Analytic Theory of Continued Fractions*, D. Van Nostrand Company, Inc., New York, N. Y., 1948. MR 10,32d

[Wil95] A. J. Wiles, *Modular elliptic curves and Fermat's last theorem*, Ann. of Math. (2) **141** (1995), no. 3, 443–551.