An (algebraic) introduction to Number Theory
Fall 2024

Kimball Martin

Revised: November 25, 2024

Contents

Preface

These are notes for MATH 4313, Introduction to Number Theory, at the University of Oklahoma in Fall 2024, and are an updated version of my notes for this course from Fall 2017. The current version of these notes should (at least for the near future) be found at the course page:

http://www.math.ou.edu/~kmartin/intro-nt/

I also taught a similar course in Fall 2009, based on the beautiful book *Elements of Number Theory*, by John Stillwell [Sti03]. While I am taking a somewhat different approach now, these notes were strongly influenced by Stillwell's presentation, as well as numerous sources that I learned from as a student.

Officially the prerequisites for the course are our Discrete Mathematics course (which is essentially an intro to proofs course) and our Linear Algebra course. We will not require the use of linear algebra in any serious way here (there will be a small amount of solving linear equations), but the ways of thinking about mathematics abstractly and axiomatically learned there will certainly be helpful, and I will make several references to things learned in Linear Algebra for the sake of comparison. We will assume the student is comfortable with reading and writing proofs, as well as the modern abstract approach to mathematics (definitions, theorems, proofs, sets, functions, equivalence relations, ...).

Number theory is a vast subject, and this course will aim to hit some of the most important topics in elementary number theory (modular arithmetic, sums of squares, quadratic reciprocity, Pell's equation, ...), but with a bent towards algebraic number theory (we'll use terminology from abstract algebra like rings and fields to talk about various examples like the Gaussian integers, though we'll avoid building up the general theory of rings and fields properly like one would in an abstract algebra course). Part of the reason for this algebraic bent is that many questions one can answer with purely "elementary" techniques are better understood from a more abstract, algebraic perspective. Time permitting, we'll also take detours into fun topics like Fibonacci numbers and continued fractions, and discuss the Riemann zeta function and distribution of prime numbers at the end of the course.[1] We'll say some more about some of these topics in the introduction.

Some pedagogical remarks: Often courses in number theory will start with easier material and build up to harder (or at least more abstract) material. This will not be our approach. We'll interleave elementary and abstract aspects, by taking a *gentle* "abstract

[1] In Fall 2017, we had a bit of time to get into continued fractions, and there's a brief section at the end of Chapter 5 involving Fibonacci numbers, but we didn't cover the Riemann zeta function. I hope to at least briefly get to the Riemann zeta function this time.

first" approach. There are a several reasons for this: (1) this is an upper-level undergraduate course, so we should attempt a somewhat serious treatment of number theory (it should not be *too* elementary); (2) given the prerequisites for this course, and the population (mostly math/CS majors), we expect the students to be able to digest abstract mathematics from the beginning; (3) standard treatments of elementary number theory make it hard to appreciate the import of basic results such as the existence and uniqueness of prime factorization of natural numbers—this is why we introduce more general number systems first so one can see how these familiar properties can fail elsewhere; (4) while I generally prefer introducing elementary situations before more abstract ones, the presentation is made more efficient by taking an "abstract first" approach; (5) by spreading out the abstract ideas throughout the course, rather than building them all up quickly at the end, I hope they will be easier to absorb; (6) (abstract) algebra is one of three main pillars of modern pure mathematics (the others being analysis and geometry/topology), and thus should be a part of any math major's training. For this approach to work, it is crucial that the students are sufficiently mathematically mature, and that the presentation is sufficiently down to earth (which is to say, the instructor and the students have to meet in the middle at some common starting ground). I hope that my expectations for students in the course is close enough to the reality that this presentation works well.

Beginning students may question the need for the abstraction—specifically algebra—that we introduce to consider what seem to be quite elementary questions about arithmetic. But we are not pursuing abstraction for abstraction's sake. The point of learning this algebra is that it will provide a lens through which we may better perceive the *structure* of arithmetic. Number theory is especially famous for having lots of elementary-to-state problems which are incredibly difficult to solve (and many remain still unsolved, as we will see in the introduction). The structure of arithmetic (e.g., prime numbers) turns out to be quite subtle, and tools from algebra (which in fact largely originated from studies into number theory) provide the best ways we have found to understand many things about numbers.

If you find errors in these notes, or have other comments/suggestions to improve them, please email me.

Introduction

Basic Terminology:

The **natural numbers** are $1, 2, 3, \ldots$.

The **integers** are $\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$.

Primes are natural numbers which have precisely 2 factors: 1 and itself; i.e., $2, 3, 5, 7, 11, 13, \ldots$. (Note for technical reasons 1 is typically excluded.)

0.1 The dictionary answer

What is number theory?

It is usually defined as the study of the integer solutions to polynomial equations with integer coefficients (called **Diophantine equations**). Some examples are $x^2 + y^2 = z^2$, $3x - 5y = 7$, $y^2 = x^3 + 12x + 5$ and $x^2 + y^2 + z^2 + w^2 = 10$. You may recognize the first equation as the Pythagorean theorem (variables suitably interpreted). In other words, the question "what are the integer solutions to $x^2 + y^2 = z^2$" is equivalent to asking what are all the integral Pythagorean triples, i.e., what are the possibilities for right-angled triangles with integral length sides. It is easy to find some—you probably remember from high school that $x = 3, y = 4, z = 5$ or $x = 5, y = 12, z = 13$ work—but how to determine all (integral) solutions is a more advanced problem.

An elegant way to solve this problem is through the use of complex numbers. In particular, define the **Gaussian integers** to be the set of numbers of the form $a + bi$ where a and b are integers and $i = \sqrt{-1}$. Thinking in terms of Gaussian integers we can factor the left hand side of the equation $x^2 + y^2 = z^2$ to get

$$\alpha\beta = (x + iy)(x - iy) = z^2.$$

Here $\alpha = x + iy$ and $\beta = x - iy$ are by definition Gaussian integers. Just like integers can be factored into primes, the Gaussian integer z^2 (which is also an integer) can be factored into what are called *Gaussian primes*, and this can be used to determine the possibilities for $\alpha = x + iy$ and $\beta = x - iy$, and hence the possibilites for x and y.

It may be helpful to illustrate the idea of using prime factorization in a simpler context. Suppose you want to find the solutions $mn = 30$ (with m, n positive integers). The prime factorization of 30 is $30 = 2 \cdot 3 \cdot 5$, so we can list all possible solutions as

$$30 = 1 \cdot 30 = 30 \cdot 1 = 2 \cdot 15 = 15 \cdot 2 = 6 \cdot 5 = 5 \cdot 6 = 10 \cdot 3 = 3 \cdot 10.$$

The idea is that we can solve the equation $\alpha\beta = z^2$ in Gaussian integers in a similar way, which leads to the complete solution (in integers) of our original equation $x^2 + y^2 = z^2$. This

is considered an **algebraic** approach. There are also so-called **elementary** approaches to this problem, as were discovered by the ancient Greeks.

Above, I said that number theory is usually defined as the study of the integer solutions of these equations. However, it is also much more this. In fact the above Pythagorean triple example illustrates several important features pervasive through number theory:

- Number theory is arguably the **oldest** branch of mathematics, beginning with counting. For a long time, mathematics was essentially just number theory and geometry.
- As questions about integer solutions can be boiled down to problems about prime numbers, perhaps the most central topic in number theory is the study of **primes** (both the familiar notion and more generalized notions such as Gaussian primes).
- Many questions in number theory have **geometric interpretations**, just as the Pythagorean triple question is a question about right-angled triangles.
- Many questions in number theory which are very simple to state are in fact very challenging to solve. In fact, unlike a course in Calculus or Linear Algebra, where most basic questions you can ask are fairly simple to solve and the subject (at its basic levels) is thought of as a "closed book," **most** basic questions you might think to ask are **still unsolved**. This has to do with the mysterious nature of prime numbers, and the richly hidden patterns in nature and numbers.

  In many cases where a solution is found, the solution will require tools from seemingly unrelated areas of mathematics. (Or rather it's often the case is that by trying to solve these problems, new areas of mathematics are discovered. It has been said that the two driving forces within modern mathematics are Number Theory and Calculus. For instance, most of Abstract Algebra (groups, rings, fields, etc) was developed out of studying problems in Number Theory.) Moreover, the beauty of many number theory problems is that the final answer is quite simple but the solution itself requires a new kind of cleverness or way of thinking.

All of these things have made number theory the branch of mathematics that, more so than any other, has fascinated amateurs and professionals throughout the ages.

0.2 Answered with questions

Another way to answer "what is number theory" is by giving you a sample of the kinds of problems studied in number theory. I hope this will make apparent the "living" nature of number theory (i.e., that people are still actively discovering new things about it), and in particular the "easy to state, hard to solve" nature of the field mentioned above which draws many mathematicians and non-mathematicians to it. Here I will describe several interesting and well known classical problems below in the form of a quiz. Some of these have been solved long ago, some not until recently and some are still unsolved. These are very roughly ordered by flavor, and not by difficulty. For each of these, I would like you to guess which have been solved long ago, which were solved in the last century (1900's), which were solved very recently (2000's), and which are still unsolved.

Bear in mind that all of these problems are well founded. In other words, while some may seem random at first, they were well thought out in advance based largely on numerical evidence.

The quiz

All numbers are assumed to be integers in the problems below, unless stated otherwise.

(1) How many primes are there?

(2) Find a formula for the n-th prime number.

(3) Are there infinitely many primes of the form $4n + 1$?

(4) Are there infinitely many primes of the form $n^2 + 1$?

(5) Note that 3 and 5, as well as 5 and 7, 11 and 13, etc. are **twin primes**, i.e., they differ by 2. Are there infinitely many twin primes?

(6) An **arithmetic progression** is a sequence of numbers such that the difference of two successive terms is constant. For example, $3, 5, 7$ (difference 2) and $11, 17, 23, 29$ (difference 6) are arithmetic progressions of primes, of lengths 3 and 4 respectively. Are there arbitrarily long arithmetic progressions of primes?

(7) Is every even integer greater than 2 is the sum of two primes?

(8) $8 = 2^3$ and $9 = 3^2$ are consecutive numbers which are both powers (squares, cubes, fourth powers, etc.) of integers. Are there others?

(9) Start with any positive n. If it is even divide by two. If it is odd take $3n + 1$. Repeat with the new number. If repeated sufficiently many times, does one eventually get down to 1 for any initial number n?

(10) Find a simple characterization of all numbers which are sums of two squares (i.e., of the form $x^2 + y^2$).

(11) Find a simple characterization of all numbers of the form $x^2 + y^2 + 10z^2$.

(12) Find a simple characterization of all numbers which are sums of 4 squares (i.e., of the form $x^2 + y^2 + z^2 + w^2$).

(13) Find a simple characterization of all natural numbers which are sums of 2 cubes of *rational* numbers.

(14) Find a simple characterization of all natural numbers which are sums of 3 cubes of *rational* numbers.

(15) Which numbers occur as areas of right triangles whose sides are all integer lengths?

(16) Are there solutions in the positive integers to $x^n + y^n = z^n$ for $n > 2$?

(17) Given any $x > 2$, do most ($\geq 50\%$) natural numbers less than x have an odd number of prime factors?

(18) Given a Diophantine equation, devise an algorithm to determine whether it has integer solutions or not in a finite number of steps.

0.3 Solutions and non-solutions

(1) How many primes are there?

Status: Easy. Solved by Euclid (ca. 300 BC). There are infinitely many primes. However, this seemingly basic question goes much deeper than this. A more refined way of asking this is: for any x, how many primes are less than x? Conjectured in 1796 by Legendre, and proved independently exactly 100 years later by Hadamard and de la Vallée Poussin, we in fact know the asymptotic distribution of prime numbers,

$$\#\{\text{primes } \leq x\} \sim \frac{x}{\log x}.$$

This result is known as the Prime Number Theorem and was proved using complex analysis and so-called the Riemann zeta function. Since many proofs (all quite difficult, but some not requiring complex analysis) have been found, until a relatively simple proof was found in 1980 by Newman (using complex analysis). The Prime Number Theorem is only a first-order asymptotic, and the "best possible" bound on the error term ($\sqrt{x}\log(x)/(8\pi)$) is equivalent to the famous (still conjectural) **Riemann hypothesis**. All of this is a central topic in **analytic number theory**. We hope to touch on this at the end of the course.

(2) Find a formula for the n-th prime number.

Status: There is no known formula (in a sense of easily computable) to generate the prime numbers, nor is it believed that there is one (at least in a simple sense). Note that such a formula would be equivalent to an exact formula for $\pi(x)$, which is quite complicated as indicated above.

(3) Are there infinitely many primes of the form $4n+1$?

Status: Known to be yes. In fact if $p(n) = an + b$ where a and b have no common factors, then $p(n)$ is prime infinitely often. This is known as **Dirichlet's theorem on arithmetic progressions** and was proved in 1837 by Dirichlet. (The case of In the course of proving this Dirichlet developed much basic groundwork used in both **algebraic** and **analytic number theory**. Time permitting, we will treat the special case of $4n+1$, which is much easier than the general form of Dirichlet's theorem.

(4) Are there infinitely many primes of the form n^2+1?

Status: Unsolved. It is easy to see that no (non-constant) polynomial can be prime for all n. However it is not known if there exists *any* quadratic (or cubic, quartic, etc.) polynomial which gives prime values infinitely often, but it is conjectured this should be true. Aside: in 1772, Euler observed that the polynomial $p(n) = n^2 + n + 41$ gives prime numbers for all $0 \leq n < 40$, but not for $n = 40$. (Clearly $p(41)$ is not prime.)

(5) Note that 3 and 5, as well as 5 and 7, 11 and 13, etc. are **twin primes**, i.e., they differ by 2. Are there infinitely many twin primes?

Status: Still unsolved. Generally believed the answer is yes. In 1966, Chen used analytic methods to show that there are infinitely many primes p such that $p+2$ is

either prime or a product of two primes. Since I first made this quiz in 2009, there has been a quantum step forward—in 2013, Yitang Zhang (an essentially unknown mathematician lecturing in New Hampshire)[2] made a huge breakthrough showing that there is some bound K such that infinitely many pairs of primes differ by at most K. We still don't know the answer to twin primes, but Zhang's work plus later refinements say we can at least take $K \leq 246$. (This would be the twin prime conjecture if we knew we could take $K = 2$, but unfortunately the known proofs do not seem capable of bringing K down to 2.)

(6) An **arithmetic progression** is a sequence of numbers such that the difference of two successive terms is constant. For example, $3, 5, 7$ (difference 2) and $11, 17, 23, 29$ (difference 6) are arithmetic progressions of primes, of lengths 3 and 4 respectively. Are there arbitrarily long arithmetic progressions of primes?

Status: Recently solved! Yes, and this was a big theorem proved by Green and Tao in 2004 using combinatorial and analytic methods (56 pages).

(7) Is every even number greater than 2 is the sum of two primes?

Status: Unsolved, though much work has been done, and the answer is believed to be yes. This was conjectured by Goldbach in a weaker form in 1742 and refined by Euler to the present form, and now called the **(strong) Goldbach conjecture**). Much progress has been made by **analytic** methods, specificially using **sieve** techniques. In 1975, Montgomery and Vaughan showed that *most* even numbers are sums of two primes. In 1995, Ramaré show that every even number is the sum of at most six primes. Since I made this quiz in 2009, the **weak Goldbach conjecture** has been solved (2013, Helfgott, building on works of others): this says that every odd number greater than 5 is a sum of 3 primes, and is called the weak Goldbach conjecture because it is implied by the strong Goldbach conjecture (the question above). So now we know weak Golbach is true, but we still don't know strong Goldbach.

(8) $8 = 2^3$ and $9 = 3^2$ are consecutive numbers which are both powers (squares, cubes, fourth powers, etc.) of integers. Are there others?

Status: Recently solved! The answer is no. This was conjectured by Catalan in 1844 and proved by Mihailescu in 2002 using **algebraic number theory** techniques (28 pages).

(9) Start with any positive n. If it is even divide by two. If it is odd take $3n + 1$. Repeat with the new number. If repeated sufficiently many times, does one eventually get down to 1 for any initial number n?

Status: Unsolved, though much work has been done. This is called the $3n + 1$ or the **Collatz problem**, proposed by Collatz in 1937. The iterated nature of the problem makes this a part of what might be called **arithmetic dynamics**, a crossroads of dynamical systems and number theory.

[2] This is a quite remarkable story and is worth reading one of the several news/magazine articles about this, e.g. https://www.quantamagazine.org/mathematicians-team-up-on-twin-primes-conjecture-20131119/

(10) Find a simple characterization of all numbers which are sums of two squares (i.e., of the form $x^2 + y^2$).

Status: Solved in 1640 by Fermat, one of the founding fathers of modern number theory (who was in fact an amateur mathematician—his profession was a judge), though not an easy problem. The solution comes by way of solving the simpler question of which *primes* are sums of two squares. The answer is precisely 2 and the primes of the form $4n + 1$! This will be one of the main theorems we prove in this course. This question, concerning squares as it does, can be interpreted geometrically, and is a starting point for the very rich area of number theory known as **quadratic forms** (meaning expressions such as x^2+y^2, $x^2+y^2+10z^2$, etc., where all terms are quadratic).

(11) Find a simple characterization of all numbers of the form $x^2 + y^2 + 10z^2$.

Status: Unsolved, but relatively recent progress. This form is known as Ramanujan's form. Ramanujan was a famous Indian mathematician who had a mystical ability to find arithmetic relations, and remarked on this form's difficulties in 1916. In 1997, Ono and Soundararajan showed that the (still conjectural) generalized Riemann hypothesis implies the following answer: all even numbers not of the form $4^k(16m + 6)$ and all odd numbers except 3, 7, 21, 31, 33, 43, 67, 79, 87, 133, 217, 219, 223, 253, 307, 391, 679, 2719. This is a famous problem in the theory of **quadratic forms**, and Ono and Soundararajan show it is intimately related to **analytic number theory** as well as **algebraic number theory** and **geometry** via **elliptic curves**.

(12) Find a simple characterization of all numbers which are sums of 4 squares (i.e., of the form $x^2 + y^2 + z^2 + w^2$).

Status: Solved. Even though you might guess that it looks harder than Ramanujan's form because of the extra variable, it's much easier, as is the answer: all integers ≥ 0. This was proved by Lagrange in 1770, and we will use some simple techniques from **algebraic number theory** to prove this result later. (Note: this problem is also easier than the case of 3 squares: $x^2 + y^2 + z^2$ which was dealt with by Legendre and Gauss decades later.)

(13) Find a simple characterization of all natural numbers which are sums of 2 cubes of *rational* numbers.

Status: Unsolved, but very recent progress! In 1995 Villegas and Zagier showed that the theory of **elliptic curves** and **modular forms** classifies, in a simple way, which *primes* are sums of 2 cubes, under the assumption of the **Birch & Swinnerton-Dyer (BSD) conjecture**, the second most famous outstanding conjecture in number theory (the first being the Riemann hypothesis, mentioned above). Recent work of Kriz (2020) proves enough about the BSD conjecture to give an answer for primes $p \not\equiv 1 \bmod 9$. The result for primes could potentially lead to the result for all natural numbers, as in the case of the sum of 2 squares, but even this is not yet clear. (You might wonder about why I asked this question for rational numbers rather than integers—this question for integers (at least for which primes are represented) reduces to the question of what numbers are represented by the quadratic polynomial $3x^2 + 3x + 1$,

which is unlikely to have a simpler description.)[3]

(14) Find a simple characterization of all natural numbers which are sums of 3 cubes of *rational* numbers.

Status: Solved by Richmond in 1923. This question is not too hard (unlike the previous one), however this problem is much harder if we ask which numbers are sums of 3 cubes of *integers*. The smallest unknown case is whether $n = 114$ is a sum of 3 integer cubes, with the previously unknown cases of $n = 33$ and $n = 42$ having just been settled by Booker and Sutherland in 2019. On the other hand, **analytic methods** have been recently applied to show that *most* numbers are sums of 3 (integer) cubes without giving any information which ones are. As both the status of this and the previous problem indicate, while the theory of quadratic forms is very rich, the theory of **cubic forms** (polynomial expressions where each term is of degree three) is still quite mysterious.

(15) Which numbers occur as areas of right triangles whose sides are all integer lengths?

Status: Unsolved! This is known as the **congruent number problem**, which seems to go back to the ancient Greeks. Interestingly enough, in 1983 Tunnell gave an elegant solution *assuming* the same conjecture Villegas and Zagier took for granted in their work on the sum of 2 cubes, the BSD conjecture.

(16) Are there solutions in the positive integers to $x^n + y^n = z^n$ for $n > 2$?

Status: Solved! You may have heard of this. The answer's no and it's called **Fermat's Last Theorem**. Wiles, with help from Taylor, proved it in 1995 using some heavy-duty **algebraic number theory** techniques (129 pages). Until then, this was the most famous unsolved problem in number theory. This proof also involves a lot of geometry via what are called **elliptic curves** and their relation to **modular forms**, which stand at a crossroads of **algebraic** and **analytic number theory**. While it would take several years of serious study to understand the complete proof, we will try to explain the case of $n = 3$ which is not too hard using some simple algebraic number theory.[4] (The cases $n = 4$, $n = 5$ and $n = 7$ are also relatively easy.) In some sense, the difficulty in general is that more general number systems do not have the nice unique factorization property that the natural numbers do.

(17) Given any $x > 2$, do most ($\geq 50\%$) natural numbers less than x have an odd number of prime factors?

Status: Solved! In 1919, Pólya conjectured the answer is yes. Indeed, if you check this for many x, it seems to be true. However, in 1958, Haselgrove proved the answer is no, without explicitly finding a counterexample, but estimating there is a counterexample of about 362 digits. In 1980, Tanaka found that Pólya's conjecture is true for $x \leq 906,150,257$ but not for $x = 906,150,258$. The point is that there are many conjectures which have been observed numerically, but turn out to be false for really

[3] See my notes on *Sums of squares, sums of cubes, and modern number theory*: http://www.math.ou.edu/~kmartin/papers/quatcubforms.pdf for more about these problems.

[4] I didn't really have time for this the last time around, so don't get your hopes up.

large numbers. There are lots of coincidental phenomena which happen for relatively small numbers that are not true in general, and this is sometimes known as the "law of small numbers." Consequently, even if you have an incredible amount of emperical evidence for a phenomenon, you still can't be sure it's true without a proof.

(18) Given a Diophantine equation, devise an algorithm to determine whether it has integer solutions or not in a finite number of steps.

Status: Solved! Sort of. Fairly recently. In 1900, Hilbert presented a famous list of 23 problems, saying that once all of these are solved, we will know all that there is to know about mathematics. (Some are more ambitious than others, and some are rather vague: The 6th is axiomatize all of physics. The 8th was the aforementioned Riemann hypothesis together with Goldbach's conjecture. Of the 23, 6 are pure number theory, and 2 of these 6 are resolved. In total, somewhere between 10 and 13 have been resolved, depending on interpretation.) This problem was Hilbert's 10th. It was resolved in 1959 by Davis and Putnam, who showed that no such algorithm exists!

Let me remark the person(s) I attribute to solving the problem are mainly just for reference purposes. A good mathematical problem gets considered by many individuals (sometimes working together, which is much more common nowadays) and the solution evolves through the effort of many people over decades or possibly centuries. In the community, people who make important contributions are usually appropriately acknowledged, but here I only mention the person(s) who completed the solution (who do of course typically deserve a lion's share of the credit). Similarly, while I occasionally gave the number of pages for the paper with the solution to give you an idea of how much it involves, bear in mind that these paper build upon previous papers, so in some sense this is just how long the "last step" of the solution is.

0.4 Main branches of number theory

Number theory can be divided into many different branches, typically delineated by the kinds of problems studied as well as the techniques used. I think most mathematicians would agree on the following as the 3 main categories of number theory, though the actual lines between them are rather blurry. These categories are divided based on the types of methods used, rather than the types of things they study.

- **Elementary number theory.** While all of the problems stated in the quiz were stated in an "elementary" way—their statement requires no advanced mathematics—very few of them can be tackled in an elementary way. One of the main ideas here is to use the idea of divisibility and some cleverness to prove some results, which one can do for things like the infinitude of primes (Euclid's answer to #1 on the quiz), the Pythagorean triple question) or which numbers are sums of squares (#10 on the quiz). Many first courses in number theory focus on elementary number theory.
- **Algebraic number theory.** The basic idea of algebraic number theory is to use other number systems to study the integers and primes, as in the example of introducing the Gaussian integers for the Pythagorean triple question. (This problem, as well as

others, are included in both the elementary and algebraic categories because there are different ways to solve it.) We could also consider problems #3, #8, and #10 – #16 in the realm of algebraic number theory.

- **Analytic number theory.** It turns out that calculus and complex analysis are very powerful tools which can be applied to number theory problems such as the Prime Number Theorem (cf. #1, #2). This is rather striking as on the surface these subjects seem very far removed from one another, but the basic idea is to consider appropriate series for studying the problem at hand. One might say that the problems #3 – #7 are in the realm of analytic number theory, though it also plays a role in problems such as #10 – #17.

- **Arithmetic geometry.** Asking for the integer solutions to, say, $x^2 + y^2 = z^2$ is the same as asking for the rational solutions to $(x/z)^2+(y/z)^2 = 1$, i.e., the *r*ational points (points with rational coordinates) on the circle $X^2+Y^2 = 1$ (where $X = x/z, Y = y/z$). In this way, many number theory problems can be translated into questions about the rational points on curves or higher-dimensional geometric objects. Arithmetic geometry is the use of methods from (algebraic) geometry to study number theory problems, and plays a central role in problems #13 – #16 above. Unfortunately, we do not have time to get into this beautiful subject in this course.

As the methods from elementary number theory tend to be limited (and doing hard things with only elementary methods can be very tedious), most number theory research nowadays involves algebraic, analytic and/or geometric methods (often mixed with elementary methods). For example, the theory of *quadratic forms* mentioned above contains aspects of elementary, algebraic and analytic number theory.

Two of the most important tools in modern number theory, as seen in applications to to #11, #13, #15 and #16 above, are:

- **Modular (**or **automorphic) forms.** These arise at a crossroads of algebraic and analytic number theory, and are closely related to things like quadratic forms and elliptic curves, as well as hyperbolic geometry. Here at OU, our number theory research group specializes more on the algebraic side (involving *groups* and *representations*) of things—in particular Ameya Pitale and I often work on the more algebraic aspects of modular and automorphic forms, and Alan Roche works on representation associated to these objects.

- **Elliptic curves.** These are related to modular forms, and are one of the central topics in arithmetic geometry. The BSD conjecture mentioned above is one of the biggest problems in the field. Elliptic curves have applications to cryptography (one member of our CS department, Qi Cheng, has worked in this area), and there's a good chance you'll see them if you take our Applied Modern Algebra course (topics vary, but this course is usually about cryptography). I do some work with elliptic curves, especially in connection to modular forms.

That said, we will not cover modular forms or elliptic curves at all in this course—which deserve full year-long courses of their own (usually at the graduate level, though the more

elementary aspects can be taught to undergraduates—certainly elliptic curves are featured in a few undergraduate texts on number theory). Instead, we will focus on more classical aspects of number theory, as mentioned in the answers to the quiz above. I just wanted to mention them to give you a bit more of an overview of number theory, and let you know a bit about the research we do here at Oklahoma.

This course will be largely elementary number theory, with some very basic algebraic number theory mixed in from the beginning, and a dash of analytic number theory at the end (Chapter 7 on the Riemann zeta function).[5]

0.5 Postscript: an example of elementary and analytic techniques

While I partially sketched an example of some simple algebraic number theory by introducing the Gaussian integers into the Pythagorean triple question, I haven't really given you any examples of elementary or analytic number theory techniques. I will illustrate each by giving two proofs of the infinitude of primes.

Theorem. *There are infinitely many primes.*

Elementary Number Theory Proof. (Euclid, ca. 300BC; also see Section 1.1 of text) This is an example of a proof by contradiction, which you should be comfortable with. Suppose on the contrary there are only finitely many primes. Label them $p_1, p_2, \dots, p_k$. Let $n = p_1 p_2 \cdots p_k + 1$. Then n divided by p_i has remainder 1 for any $i = 1, 2 \dots k$, i.e., none of the p_i's are factors of n. This leaves two possibilities: either n itself is prime (if it has no factors besides 1 and n), or it is not. If n is prime, we have our contradiction and are done.

If n is not prime, $n = ab$ for some $1 < a, b < n$. Since no p_i is a factor of n, no p_i is a factor of a either. Now we repeat our argument for n with a: either a is prime, or not. It a is prime, we are done. If not, we apply the argument again with a smaller factor of a. Now this process must terminate in a finite number of steps (less than n), because we are working with smaller and smaller integers between 1 and n. Thus we will eventually end with a prime factor of n, contradicting the assumption that there were only finitely many primes. (This process of going down from n to a and so on is called *descent*; cf. Section 1.2.) □

Analytic Number Theory Proof. (Euler, ca. 1735) The key idea of Euler is to observe that

$$\left(1 + \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \cdots\right)\left(1 + \frac{1}{3} + \frac{1}{3^2} + \frac{1}{3^3} + \cdots\right)\left(1 + \frac{1}{5} + \frac{1}{5^2} + \frac{1}{5^3} + \cdots\right)\cdots$$
$$= 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \cdots = \sum_{n=1}^{\infty} \frac{1}{n},$$

where the product on the left is a product of the quantities

$$1 + \frac{1}{p} + \frac{1}{p^2} + \frac{1}{p^3} + \cdots$$

[5] Time permitting.

as p ranges over all primes. Note that this series is a geometric series with ratio less than 1, so it is evaluated by

$$1 + \frac{1}{p} + \frac{1}{p^2} + \frac{1}{p^3} + \cdots = \frac{1}{1 - 1/p}.$$

(If you forgot this, multiply through by the denominator of the right hand side, and the left hand side telescopes down to 1.) Hence we have

$$\sum_{n=1}^{\infty} \frac{1}{n} = \prod_{p} \frac{1}{1 - 1/p} = \infty$$

since the left hand side is the harmonic series which diverges. Since each term in the product over primes is a finite number, for this product to diverge, it must be infinite. I.e., there must be infinitely many primes! In other words, the infinitude of primes is equivalent to the divergence of the harmonic series! □

While the analytic proof may seem unnecessarily complicated (in that it involves some calculus—it is not actually longer), i) it is certainly beautiful, and ii) the basic ideas in this proof can be pushed much much further to get strong results like the Prime Number Theorem, which one can't do with Euclid's proof. We'll discuss these ideas, without giving many proofs, in Chapter 7.[6]

0.6 Notation

There is an index in the back which includes key definitions and notation defined in the text. Here are some additional conventions about notation. (Some of these comments might not make sense to you now.)

- for sets A, B, we use $A \subset B$ to mean A is a subset of B (not necessarily a proper subset)
- for sets A, B, we use $A \sqcup B$ to mean disjoint union, i.e., it means $A \cup B$ and the statement that $A \cap B = \emptyset$
- for sets A, B we use $A - B = \{a \in A : a \notin B\}$ for the set difference (we do not require $B \subset A$ to write $A - B$)
- for us, the natural numbers $\mathbb{N}$ begin at 1, not 0
- for integers a, b, we use $a \mid b$ to mean a divides b, i.e., $b = ka$ for some integer k, and we use $a \nmid b$ to mean a does not divide b
- ring means commutative ring unless stated otherwise
- p will typically denote a prime number (2, 3, 5, 7, 11, ...), or more generally a prime element of a ring

[6] I haven't actually written this yet, as we didn't get that far when I taught this before. Fingers crossed for this time!

0.7 References

If you're looking for supplementary references, there are *many, many* introductions to number theory. However, most books I know of are either more elementary than this class or more advanced (e.g., most books that do some algebraic number theory assume a course in elementary number theory first). Or they might cover a lot of the same material, but they also cover a lot more, and with a rather different presentation.

But since some students have asked about other references, here is a short list of possibilities. You can find many more by searching online or going to our library. Each reference has its own approach with its own advantages and disadvantages, so if you want supplementary presentation/material, browsing until you find something appealing is not a bad way to go.

All of the following references are free online through our library.

- *Elements of Number Theory*, by John Stillwell [Sti03]. As mentioned in the preface, I've used this before, and I think it's a great little book.[7] Of the books I know, this is perhaps the closest in content to the current course, but it's not an exact match, and I'm presenting the material in a rather different way this time.

- *Elementary Number Theory*, by GA Jones and MA Jones [JJ98]. I think this book is a nice introduction, and I was seriously contemplating using this as our text. I believe most of the main results we will prove are covered in here, though again our approach and presentation will be different.

- *Elementary Number Theory: Primes, Congruences and Secrets*, by William Stein [Ste09]. This is another book I briefly considered using for this course. It is more cryptographically and computationally oriented than what I wanted to do with this class, but probably overlaps with a little over one-half of the content of our course. If you like the bits we do related to RSA and cryptography, try here for more.

- *The Whole Truth About Whole Numbers*, by Sylvia Forman and Agnes Rash [FR15]. Disclaimer: I have not looked at this book personally, I just found it when looking to see what references we have electronic access to. It seems to be more elementary and more cryptography-oriented than what we will do, and maybe overlaps with one-half of the content of our course. Chapter 2 goes over proofs, so if you need to bone up on your proof ability, maybe this will be a good reference.

Furthermore, if you do look at some of these or other references, I'd appreciate hearing your thoughts. It will help me make recommendations/choose materials in the future.

[7] You can also see my online supplementary notes from when I taught from that book here: https://math.ou.edu/~kmartin/nti/

Chapter 1

Numbers

While number theory is about studying equations over the integers or rationals, one of the primary tools to study these is by using auxiliary number systems, such as the Gaussian integers as indicated in the introduction. A more basic type of number system you may be familiar with is modular arithmetic. For instance, a simple application of modular arithmetic is that no number of the form $4n + 3$ is a sum of two squares.

This chapter will start with the numbers you know from grade school and move on to more general number systems. Along the way, we'll introduce the terms *ring* and *field* from abstract algebra (which were in fact motivated from number theory) as a convenient tool to talk about the some of the basic ways in which number systems can differ.

1.1 Standard number systems

If we think back to the murky origins of mathematics, counting is surely at the very beginning.[1] From early on, humans could distinguish between having one apple and multiple apples, or one thing with pointy teeth coming after me versus many things with pointy teeth coming after me. Learning to count things was eminently useful, and at some point humans made the conceptual leap from "5 apples" to the abstract notion of 5. (Rabbits on the other hand, can only count to four—see Section 1.4.) We gave names to numbers, at least small numbers at first, and eventually we learned how to talk about big and small numbers. (Though some tribal languages, e.g., the Amazonian language Pirahã, are claimed to have no words for precise numbers.) This gave us the **counting numbers**, or **natural numbers**. The set of natural numbers are

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

The basic arithmetic operations we can do on $\mathbb{N}$ are addition and multiplication. **Number theory**, also sometimes referred to as **arithmetic**, is really about understanding numbers with respect to these operations, as opposed to thinking about individual numbers in and of themselves.

[1] Though recognition of shapes, and thus some sort of notion of geometry, also must have occurred at the beginning—I am making no claims as to which came first, or if they arose at about the same time.

We note that it is clear to us now that there are infinitely many natural numbers. (Proof: Suppose not. Then there is a largest number N. But then $N+1$ is bigger, a contradiction.) However it may not have always been obvious, and there are philosophical positions (not widely held, admittedly) positing that really large numbers do not actually "exist" in some sense.[2]

Likely, natural numbers were first represented as a series of ticks, so 3 was represented by |||. Various numeral systems (e.g., Arabic, Roman) have been developed, with so-called positional systems coming to dominate. Note that using series of ticks to represent numbers allows us to represent any natural number we have the resources to record, but is very inconvenient for representing large numbers. (However, the rules for addition and multiplication are quite simple in this system.)

The type of numeral system we most commonly use today is known as a positional system. It requires having a zero place holder, and is most convenient to introduce after the next big conceptual leap in numbers: zero. The numbers consisting of natural numbers and zero are known as the **whole numbers**[3], or the non-negative integers. We will denote this set as

$$\mathbb{Z}_{\geq 0} = \mathbb{N} \cup \{0\} = \{0, 1, 2, 3, \ldots\}.$$

Now we can define a **base b positional system** as follows. Let $X_0, \ldots, X_{b-1}$ be distinct symbols representing the numbers $0, \ldots, b-1$. Let a_i denote a number in $0, 1, \ldots, b-1$ for $0 \leq i \leq m$. Then we equate a string of symbols X_{a_i} with a whole number as follows:

$$X_{a_m} X_{a_{m-1}} \cdots X_{a_1} X_{a_0} = a_m b^m + a_{m-1} b^{m-1} + \cdots + a_1 b + a_0.$$

This definition looks more complicated than it is, which the following examples should make clear.

If $2 \leq b \leq 10$, we typically just use the standard Arabic numeral for j as our symbols X_j. So with $b = 10$, our 10 symbols are the usual $0, 1, \ldots, 9$. Then the above just becomes the **decimal system** that you are familiar with from grade school. For instance,

$$7083 = 7 \cdot 10^3 + 0 \cdot 10^2 + 8 \cdot 10 + 3.$$

Besides the decimal system, the second most common positional system now is most likely **binary**, which is base 2, thanks largely to its uses in computer science. For instance, the first few numbers in binary are

[2] Edward Nelson, a math professor at Princeton until 2013, was a notable skeptic of the logical consistency of the infinitude of numbers. Other respected mathematicians have also expressed skepticism about the infinitude of numbers (e.g., https://arxiv.org/abs/math/0605779) though this is certainly a minority position beyond elementary school. We'll touch on this again in Section 1.4.

[3] Some authors include zero in the natural numbers, but in the US the definition I gave for $\mathbb{N}$ is standard.

binary		decimal
0	0	0
1	1	1
10	$1 \cdot 2 + 0$	2
11	$1 \cdot 2 + 1$	3
100	$1 \cdot 2^2 + 0 \cdot 2 + 0$	4
101	$1 \cdot 2^2 + 0 \cdot 2 + 1$	5
110	$1 \cdot 2^2 + 1 \cdot 2 + 0$	6
111	$1 \cdot 2^2 + 1 \cdot 2 + 1$	7

The way we have define a base b positional system works for any integer $b \geq 2$. For $b = 1$, it should be defined slightly differently (not using zero), and amounts to the system of using n tick marks to represent the number $n \in \mathbb{N}$. Base 1 is also called **unary**. Note that one cannot represent 0 in unary, at least not unambiguously. (Zero in unary would be represented by no tick marks, but then there's no way to distinguish between 0 and blank space on a page.)

Note there is no abstract mathematical reason why decimal is natural or better to use than other positional systems (and certainly no mathematical reason why we use Arabic numerals), but this is rather a function of our biology, having (in most cases) 10 digits on our hands.

Exercise 1.1.1. The base b positional system with $b = 16$ and the symbols $0, \ldots, 9$, A, B, ... F representing $0, \ldots, 15$ is called **hexadecimal**, and is also used in computer science. Write the decimal numbers 16, 32, and 200 in hexadecimal.

Exercise 1.1.2. Let s_n be a string of n 1's, which we view as representing a number in binary. What is the binary representation for $s_n + 1$? (Prove your answer is correct.)

Exercise 1.1.3. Compute 201+112 in base 3. (Here 201 and 112 are base 3 representations of the number, and you should write your answer in base 3 also). Then state what all of these numbers are in decimal.

The basis for much of elementary number theory is the following.

Theorem 1.1.1 (Fundamental theorem of arithmetic)**.** *Let $n > 1$ be a natural number. Then n factors into a product of prime numbers. Moreover, this factorization is unique up to reordering, i.e., if*

$$n = p_1 p_2 \cdots p_r = q_1 q_2 \cdots q_s,$$

where the p_i's and q_j' are primes, and are ordered so that

$$p_1 \leq p_2 \leq \cdots \leq p_r, \quad q_1 \leq q_2 \leq \cdots \leq q_s,$$

then $r = s$ and $p_i = q_i$ for each $1 \leq i \leq r$.

Note the fundamental theorem of arithmetic consists of two parts: the existence of a prime factorization (when $n \neq 1$), and the uniqueness of this prime factorization.

Here, as in the introduction, a prime number is a natural number $p > 1$ with exactly two factors (divisors), 1 and p.[4] This should be familiar to you, but we prove the existence of a prime factorization below in Proposition 1.1.3 and will give a proof of the uniqueness later in Chapter 2. The proof (particularly for uniqueness) is not entirely trivial, and we will see that the proof works for some number systems, but not others, as in fact many number systems we will consider do not have prime factorization. (Also, the definition of prime for other number systems will not be the same as the one we gave for natural numbers, though it happens to be equivalent to the familiar one we gave above in the case of $\mathbb{N}$.)

In the above statement, note that many of the p_i's may be the same in the factorization $p_1 p_2 \cdots p_r$. E.g., $500 = 2 \cdot 2 \cdot 5 \cdot 5 \cdot 5$. However, for many number theory arguments, it's convenient to group all of the equal primes together, i.e., to write

$$n = p_1^{e_1} p_2^{e_2} \cdots p_k^{e_k}, \tag{1.1.1}$$

where each p_i is prime and $p_i \neq p_j$ for $i \neq j$ $(1 \leq i, j \leq k)$. E.g., $500 = 2^2 \cdot 5^3$. We will call a factorization of this form the **prime-power factorization** of n. (We use the article "the" even though there is technically a dependence on the ordering of the p_i's. If the order matters for an argument, we will specify the order at the time.) When we talk about the **prime factorization** of n, by default we will mean the prime-power factorization, but we may also use it for factorizations of the form $p_1 \cdots p_r$, or even $p_1^{e_1} \cdots p_r^{e_r}$, where not all p_i's need be distinct. In the latter event, we will specify that not all p_i's need be distinct if it is not clear.

In terms of the prime-power factorization, uniqueness of prime factorization means that if

$$n = p_1^{e_1} \cdots p_r^{e_r} = q_1^{f_s} \cdots q_s^{f_s},$$

then $r = s$, and after relabeling q_j's if necessary, we have $p_i = q_i$ and $e_i = f_i$ for all $1 \leq i \leq r$.

1.1.1 An axiomatic approach

At extremes, there are two kinds of approaches to mathematics: an *intuitionistic approach*, and a *formalistic approach*. The intuitionistic approach goes back to the very beginnings of mathematics, and represents our natural way of learning and thinking about things. However, like science, many earlier "results" found this way turned out later to be incorrect, and we needed to revise our understanding to get closer and closer to the truth. The formalistic approach, going back at least to Euclid's approach to geometry, is rooted in logic, and attempts to make mathematics 100% correct, given starting axioms and logical rules of inference to reason about them with. The axioms and rules of inference represent things we take for granted about reality, though we cannot ever (formally, i.e. with 100% certainty) prove that they provide an accurate representation of reality.

In practice, mathematicians typically work somewhere in between a completely intuitionistic approach and a completely formalistic approach. Though many abstract math

[4] This theorem is perhaps the first reason why we do not allow 1 to be prime: if we did, prime factorizations would not be, e.g., $2 \cdot 3$ and $1 \cdot 1 \cdot 2 \cdot 3$ would both be prime factorizations of 6.

classes may seem very formal to you, outside of serious logic courses, they are typically quite far from complete formality—we rarely justify all of our reasoning all the way down to the starting axioms (indeed, we rarely give a precise definition of a "set") and valid rules of inference—this would be far too tedious, as well as rob us of both the power and the beauty of intuition and creativity.

There are various axiomatic models of the natural numbers, with the most famous being **Peano's axioms** from 1889, which I'll summarize here already assuming set theory, just to give you an quick idea. Peano's axioms declare a set $\mathbb{N}$, called the set of natural numbers, which satisfies:

(1) There is an object $1 \in \mathbb{N}$.

(2) There is a function $\mathrm{succ} : \mathbb{N} \to \mathbb{N}$ called the **successor function**. (Think: $\mathrm{succ}(n) = n + 1$ is the next number after n)

(3) succ is an injection, i.e., $\mathrm{succ}(m) = \mathrm{succ}(n) \implies m = n$.

(4) There is no $n \in \mathbb{N}$ such that $\mathrm{succ}(n) = 1$.

(5) **[induction axiom]** If $S \subset \mathbb{N}$ such that $1 \in S$ and $\mathrm{succ}(S) \subset S$ (i.e., $\mathrm{succ}(n) \in S$ for all $n \in S$), then $S = \mathbb{N}$.

We remark the axiomatic approach avoids the question of what a number really is (which is perhaps best understood by us intuitively anyway). Technically, we should say that the natural numbers $\mathbb{N}$ are just a **model** for the Peano axioms, as the Peano axioms do not uniquely characterize $\mathbb{N}$, but we will not deliberate on these subtleties here.[5] This technicality aside, we can think of $\mathbb{N}$ formally as follows: 1 is just 1, 2 is defined to be $\mathrm{succ}(1)$, 3 is defined to be $\mathrm{succ}(2) = \mathrm{succ}(\mathrm{succ}(1))$, and so on. From these axioms (and a reasonable logical system), one can define addition and multiplication of natural numbers and show they satisfy the usual properties (commutative, distributive, etc). We will not do this, and take a reasonable model of $\mathbb{N}$ for granted, but it is good to be aware of the axiomatic treatment and that these things can be made more formal if desired. The following exercise suggests how to proceed.

> **Exercise 1.1.4.** In Peano's model, for $m, n \in \mathbb{N}$, define $m + n$ to be the m-fold successor of n (e.g., $2 + n = \mathrm{succ}(\mathrm{succ}(n))$). Using Peano's axioms, prove that with this definition, $2 + 3 = 3 + 2$.

It's often helpful to think in terms of pictures, so it's useful to have a visual representation of number systems. We can view $\mathbb{N}$ and $\mathbb{Z}$[6] as in Fig. 1.1.1. In terms of the picture for $\mathbb{N}$,

[5] The most basic issue is that there is nothing in the Peano axioms saying each number in the system has to be "finite." And this is not something we can easily put in, because how do you define finite? You need to use something like $\mathbb{N}$ already, and this would lead to circular reasoning. There are other number systems that satisfy Peano's axioms, but such things are more suitable for a course in logic rather than number theory.

[6] $\mathbb{Z}$ denotes the integers, as you should have learned before, though I will define $\mathbb{Z}$ again at the beginning of the next section.

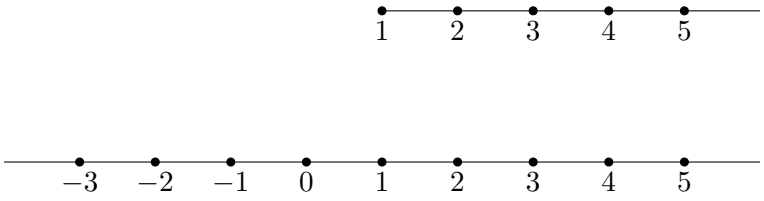


Figure 1.1.1: Visualizing $\mathbb{N}$ and $\mathbb{Z}$

you can think of the first two Peano axioms as saying: 1) draw a dot, and label it 1; and 2) for each dot you draw, you must draw another dot to to the right. Then the picture for $\mathbb{Z}$ suggest that one can extend Peano's axioms to $\mathbb{Z}$ by introducing a rule that "says for each dot you draw, you must draw another dot to the left," or more formally what one would call a predecessor function.

Exercise 1.1.5. Explain what the latter 3 Peano axioms mean in terms of the picture.

Exercise 1.1.6. Try to formulate an analogue of Peano's axioms for $\mathbb{Z}$.

The induction axiom guarantees that mathematical induction is a valid way to prove properties of $\mathbb{N}$. (This should be intuitively clear from Exercise 1.1.5.) It has an important consequence for us:

Proposition 1.1.2 (**Descent principle**). *Any strictly decreasing sequence of natural numbers is finite.*

The descent principle is commonly stated in slightly different terms, and called the **least integer principle**. This principle commonly arises in proofs in number theory, with what is called the **(Fermat's) method of descent** (or **infinite descent**) (though the method goes back at least to Euclid), which is really an induction proof in disguise. We will often just call this **descent** for short.

The basic idea with the method of descent is to reduce a problem to "minimal cases." For instance, say we want to prove a statement $S_n : 1 + 2 + \cdots + n = \frac{n(n+1)}{2}$ for all $n \in \mathbb{N}$. It is not hard to show that S_n holds if S_{n-1} does. (Logically this is written $S_n \Longleftarrow S_{n-1}$ or $S_{n-1} \implies S_n$ with arrows pointing from $n-1$ to n—it is not $S_n \implies S_{n-1}$). In other words, for any n, we can reduce case n to the $n-1$ case, and therefore the $n-2$ case, and so on. By the descent principle, this process must terminate, and we eventually end at the $n = 1$ (or $n = 0$, if you prefer) case, which is trivial to check. Try writing this up carefully yourself:

Exercise 1.1.7. Prove that $1 + 2 + \cdots + n = \frac{n(n+1)}{2}$ for all $n \in \mathbb{N}$ using descent.

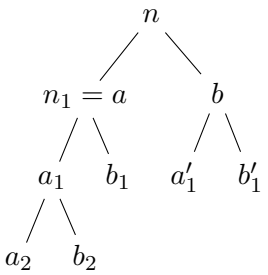
Of course in the above example, there is no real advantage to descent over induction. But it many situations, descent provides a way to think about a problem that may be more intuitive than induction. Here is one that is of importance for us. (I'll leave it to you whether or not you think this is more intuitive than induction—see the exercise below.)

Proposition 1.1.3 (Existence of prime factorization)**.** *Let $n \in \mathbb{N}$ with $n \neq 1$. Then there exists a sequence of (not necessarily distinct) prime numbers $p_1, \dots, p_r$ such that $n = p_1 \cdots p_r$.*

Proof. Either n is prime or not. If so $n = p$ is a prime factorization, and we are done. Therefore, suppose n is not prime, i.e., it has a factor a which is neither 1 nor n, so we can write $n = ab$ for some $a, b \in \mathbb{N}$. Since a is not 1 or n, neither is b. Now note it suffices to prove a and b have prime factorizations, say $a = p_1 \cdots p_s$ and $b = p_{s+1} \cdots p_r$, for then $n = p_1 \cdots p_r$. (Put another way, it suffices to prove all natural numbers $< n$ greater than 1 have a prime factorization.)

So we simply repeat the above arguments for a and b. Let's just consider $n_1 = a$. Either n_1 is prime or not. If n_1 is prime, we are done. If not, we can factor $n_1 = a_1 b_1$ where $1 < a_1, b_1 < n_1$. This reduces the problems to proving the existence of prime factorization for numbers less than $n_1 < n$.

It may be helpful to think of this argument as constructing a "factorization tree" as follows:



Here we keep dissecting the tree as long as the numbers we get are not prime, so when we are done all the leaves (nodes with nothing below them) at the bottom of the tree must be prime. In particular, if the above picture represents a completed factorization tree, it represents the prime factorization $n = a_2 b_2 b_1 a_1' b_1'$.

To finish the proof, we have to prove that along any path we take in this factorization tree (going from top to bottom in some way), we eventually stop at a prime number. That is, this process of breaking up factors into smaller factors can't go on forever. (In computer science lingo, we need to show the above algorithm for constructing the factorization tree eventually terminates.) Indeed, since at each stage in this recursive argument, we are getting smaller and smaller numbers, this process eventually arrives at a prime factorization by the principle of descent. (If not, there would be some infinite sequence of natural numbers $n > n_1 > n_2 > n_3 > \cdots$ of successive non-prime factors, which is impossible by descent.) □

Exercise 1.1.8. Rewrite the above proof of existence of prime factorization using strong induction instead of descent.

We emphasize that the above argument does not prove that each n has only one prime factorization (up to reordering). Indeed, we will see below the same argument applies in situations where one does not have unique factorization. The issue is that there can be many ways one can to factor any $n_i = a_i b_i$, so there are many possible factorization trees for n. We will need another argument to guarantee that the leaves (primes) of all factorization trees for n are the same. We will see this in Chapter 2.

1.1.2 Beyond counting

In our daily reckonings besides addition, its inverse, subtraction, is also very useful. But since one cannot subtract arbitrary natural or whole numbers and get a whole number, one needs to introduce the notion of negative numbers. This was surely a great leap in abstraction, and really required an abstract notion of a number quite divorced from representing a physical number of objects. This extends the whole numbers we know to give the **integers**:

$$\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}.$$

(The $\mathbb{Z}$ is for *Zahlen*, which means integer in German.)

Just like we need to do the opposite of addition sometimes, we also need to do the opposite of multiplication, which is division. We cannot divide arbitrary (nonzero) integers and remain in the set of integers. Rather division leads us to the **rational numbers**

$$\mathbb{Q} = \left\{ \frac{a}{b} : a \in \mathbb{Z}, b \in \mathbb{N} \right\}.$$

(The $\mathbb{Q}$ is for quotients.) Note that unlike for $\mathbb{N}, \mathbb{Z}_{\geq 0}$ and $\mathbb{Z}$ where our standard representation of a number is unique, the above representation of rationals is not unique, e.g., $\frac{1}{2} = \frac{5}{10}$. To get uniqueness of a representation $\frac{a}{b}$, we have to assume $\frac{a}{b}$ is in **reduced form**, i.e., $a \in \mathbb{Z}$ and $b \in \mathbb{N}$ where a and b have no common prime factors. (When $a = 0$, reduced form should be interpreted as taking $b = 1$. Another way to describe reduced form is to say that the denominator is as small as possible.)

Recall that a relation $\sim$ on a set A is an **equivalence relation** if (i) $a \sim a$ for all $a \in A$ (reflexivity), (ii) $a \sim b$ implies $b \sim a$ for all $a, b \in A$ (symmetry), and (iii) $a \sim b$ and $b \sim c$ implies $a \sim c$ for all $a, b, c \in A$ (transitivity). Recall also that if $\sim$ is an equivalence relation on A, then $\sim$ partitions A into subsets called **equivalence classes**, which consist of all elements of A which are equivalent to each other.

Exercise 1.1.9. (i) Show $(a, b) \sim (c, d)$ when $ad = bc$ defines an equivalence relation on $\mathbb{Z} \times \mathbb{N}$.[7]

(ii) Prove that $\mathbb{Q}$ is in natural bijection with the $\mathbb{Z} \times \mathbb{N}/\sim$, the equivalence classes of $\mathbb{Z} \times \mathbb{N}$ for the equivalence relation in (i).

There are two other major number systems you know about beyond the above 4 that number theory is most directly concerned with. Already by the time of the ancient Greeks, it was known that certain geometrical quantities (e.g., $\sqrt{2}$, which is the hypotenuse of an right triangle with other side lengths both 1) are not rational numbers. In order to describe

arbitrary geometric quantities, we have the **real numbers** $\mathbb{R}$.[8] (A proper definition of $\mathbb{R}$ is somewhat complicated—a couple of traditional ways are using *Cauchy sequences* and *Dedekind cuts*, which you may learn in an analysis class. Similar to $\mathbb{Q}$, the real numbers $\mathbb{R}$ also have the issue that the standard decimal representation is not unique, e.g., $0.999\ldots = \frac{9}{9} = 1.000\ldots$.)

However, to do algebra in general, $\mathbb{R}$ is not quite sufficient, and one considers the **complex numbers**

$$\mathbb{C} = \{x + iy : x, y \in \mathbb{R}\},$$

where i is defined to be a square root of -1. (There are two such square roots, with $-i$, being the other.) The representation of a complex number as $x + iy$ with x, y real is unique (given representations for x, y).

The **fundamental theorem of algebra** says that any polynomial $c_n z^n + c_{n-1} z^{n-1} + \cdots + c_1 z + c_0$ of degree $n \geq 1$ (so $c_n \neq 0$) with factors into n linear polynomials over $\mathbb{C}$:

$$c_n z^n + c_{n-1} z^{n-1} + \cdots + c_1 z + c_0 = c_n (z - a_1)(z - a_2) \cdots (z - a_n),$$

for some $a_1, \ldots, a_n \in \mathbb{C}$. What this means is that if we have any single variable polynomial equation, we can always solve it over $\mathbb{C}$ (and in fact say how many solutions their are, counting multiplicity). For instance, solutions to the equation $z^2 = -1$ corresponds to roots of the polynomial $z^2 + 1 = (z - i)(z + i)$, and there are two solutions: $z = \pm i$. As mentioned in the introduction, this property of $\mathbb{C}$ is supremely important in number theory. We won't prove this theorem (places you might see a proof: algebra, complex analysis, or topology classes), the simplest (nontrivial) case is an easy exercise:

Exercise 1.1.10. Prove the fundamental theorem of algebra for $n = 2$.

Exercise 1.1.11. Find the roots of the polynomial $z^2 + z + 1$. Show they satisfy $z^3 = 1$.

A summary of the main features/differences of these number systems is in Table 1.1. The "operations" column indicates what operations we can do to pairs of numbers in the given system and get back a number within the same system (excluding division by 0 in the case of $\div$).

Remark 1.1.4. The phrase "standard number systems" is not itself standard. (I'm not even sure how widespread "number system" is, to be honest.) I just mean it to refer the above 6 number systems which you should be familiar with from primary and secondary school.

1.2 Rings and fields

In this section, we'll introduce some mathematical language for talking about different types of number systems. This section may seem rather abstract, but really it's just about having precise language to talk about two important kinds of number systems.

[8] Perhaps the most important thing about $\mathbb{R}$ is that limits exist—e.g., a bounded monotone sequence of real numbers is a real number, but this is not true for $\mathbb{Q}$. This lets us do calculus/analysis on $\mathbb{R}$, but this limit property, known as *completeness*, will not play a major role in our present course.

number system	symbol	operations	remarks
natural numbers	$\mathbb{N}$	$+, \times$	
whole numbers	$\mathbb{Z}_{\geq 0}$	$+, \times$	has 0
integers	$\mathbb{Z}$	$+, -, \times$	
rationals	$\mathbb{Q}$	$+, -, \times, \div$	
reals	$\mathbb{R}$	$+, -, \times, \div$	can be ordered (i.e., have $<$ and $>$)
complex numbers	$\mathbb{C}$	$+, -, \times, \div$	can take roots of polynomials

Table 1.1: Standard number systems

1.2.1 Binary operations

First we begin with a fundamental mathematical definition, which you may have seen in an earlier course.

Definition 1.2.1. *Let S be a set. A* **binary operation** *on S is a map $* : S \times S \to S$. That is, it is a way of assigning to each pair of elements $(x, y) \in S \times S$ a uniquely defined element $x * y \in S$.*[9]

The most fundamental examples of binary operations are $+$ and $\times$ on $\mathbb{N}$ or $\mathbb{Z}$. However, since we'll also work with other binary operations and sets, let's first think about general binary operations a bit.

If S is a finite set, say with cardinality n, then a binary operation on S can be specified by an operation table, e.g., if $S = \{a, b, c\}$, then one such table is

$*$	a	b	c
a	a	a	b
b	c	b	a
c	c	c	c

We read this as defining an operation $*$ by letting $x * y$ be the entry corresponding to row x and column y in the table. For instance, with the above operation, $a * a = a$, $a * b = a$ and $b * a = c$.

The number of possible operation tables is the number of functions from $S \times S$, a set of size n^2, to S, a set of size n. More concretely, to make an operation table, we have n^2 entries to fill in (once we label the rows and columns by elements of S, in some order we choose), each of which can be one of n possible elements. Hence there are a total of n^{n^2} operation tables (once we fix a labelling for rows and columns), i.e., n^{n^2} binary operations, on a set of size n.

Note that an arbitrary binary operation does not possess any special properties. For instance $x * y \neq y * x$ in general, as the above example shows.

If $*$ is a binary operation on a set S, we say $*$ is **commutative** if

$$x * y = y * x \quad \text{for all } x, y \in S. \tag{1.2.1}$$

[9] Fun fact: a set with a binary operation is called a *magma*.

Commutativity is perhaps the most basic property you might expect an operation to have, and many operations we are familiar with such as $+$ and $\times$ have this property.

Another basic property you might want an operation to have is associativity: we say a binary operation $*$ on a set S is **associative** if

$$(x * y) * z = x * (y * z) \quad \text{for all } x, y \in S. \tag{1.2.2}$$

I.e., associativity means the order of operations does not matter. (Actually, since several natural operations are associative but not commutative—we'll see a couple of examples below—associativity might be considered more "basic" than commutativity.) Again, $+$ and $\times$ have this property, and a common mistake for students is to assume all operations do.

Another basic property a binary operation $*$ on a set S can have is the existence of an **identity (element)**. This is an element $e \in S$ such that

$$e * x = x * e = x, \quad \text{for all } x \in S. \tag{1.2.3}$$

If $S = \mathbb{Z}$ and our operation is $+$ (resp. $\times$) then 0 (resp. 1) is an identity element.

Exercise 1.2.1. Write down the operation tables for all binary operations on $S = \{a, b\}$. Which are commutative? Which have an identity element?

Exercise 1.2.2. Let $S = \{T, F\}$. Interpret the logical operations 'and' ($\wedge$), 'or' ($\vee$) and 'xor' ($\oplus$) as binary operations on S.

Exercise 1.2.3. Let S be a set of size n. How many commutative binary operations are there on S?

Exercise 1.2.4. Prove that a binary operation $*$ on a set S has at most one identity element. (*Hint:* Try contradiction.)

Exercise 1.2.5. Let S be a set of size n. How many binary operations on S have an identity element?

Intuitively, a binary operation on a set S is simply a way of combining two elements of S to get a new element. If the operation is commutative, this means it doesn't matter in what order we combine our elements, but if the operation is non-commutative, it does. (Non-commutative means that the operation is not commutative, i.e., $x * y \neq y * x$ for some $x, y \in S$. It does not mean $x * y \neq y * x$ for all $x, y \in S$.) If the operation has an identity e, this means combining e with any other element x yields x again. You can think of this as saying combining with e doesn't do anything.

Now let's go through the 4 most basic arithmetic operations on the various standard number systems.

Example 1.2.1. Addition ($+$) is a commutative, associative binary operation on any of the following sets: $\mathbb{N}$, $\mathbb{Z}_{\geq 0}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$. Furthermore, on any of these sets except $\mathbb{N}$, $+$ has an identity element, 0, which we also call the *additive identity*.

Example 1.2.2. Subtraction ($-$) is a binary operation on any of the following sets: $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$. Note that subtraction is neither commutative nor associative, e.g. $1-0 \neq 0-1$ and $(1-1)-1 \neq 1-(1-1)$. Moreover, subtraction it is not a binary operation on $\mathbb{N}$ or $\mathbb{Z}_{\geq 0}$ as $1-2$ does not give another element of $\mathbb{N}$ or $\mathbb{Z}_{\geq 0}$.

Exercise 1.2.6. Show that subtraction on $\mathbb{Z}$ does not have an identity element.

Example 1.2.3. Multiplication ($\times$ or $\cdot$) is a commutative, associative binary operation on any of the following sets: $\mathbb{N}$, $\mathbb{Z}_{\geq 0}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$. Furthermore, on all of these sets, $\times$ has an identity element, 1, which we also call the multiplicative identity.

Example 1.2.4. Division ($\div$ or $/$) is *not* a binary operation on any of the following sets: $\mathbb{N}$, $\mathbb{Z}_{\geq 0}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$. The issue with $\mathbb{N}$, $\mathbb{Z}_{\geq 0}$ or $\mathbb{Z}$, is that we can divide two natural numbers or integers and not get an integer, e.g., $1/2 \notin \mathbb{Z}$. The issue with $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ is that division by zero is undefined, i.e., $1/0$ is not a well-defined element of $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$. (This is also an issue for $\mathbb{Z}_{\geq 0}$ and $\mathbb{Z}$, so there are actually two reasons why the division is not a binary operation for whole numbers and integers.) We remark that we could *define* division by 0 (e.g., declare $x/0 = 0$ for all x) to make it a binary operation on $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$, however we don't want to because this would screw up properties we want division to have, such as $b \cdot \frac{a}{b} = a$ for all $\frac{a}{b} \in \mathbb{Q}$.

However, there is another way to make division into a binary operation that is better. Let $\mathbb{Q}^\times$ (resp. $\mathbb{R}^\times$, resp. $\mathbb{C}^\times$) denote the set of nonzero elements of $\mathbb{Q}$ (resp. $\mathbb{R}$, resp. $\mathbb{C}$). Then division is a binary operation on any of the sets: $\mathbb{Q}^\times$, $\mathbb{R}^\times$ and $\mathbb{C}^\times$. Like subtraction, it is neither commutative nor associative, and does not possess an identity element.

Just to point out that binary operations abound in mathematics, I'll give a couple more examples you're probably familiar with.

Example 1.2.5. Let $M_n(\mathbb{R})$ be the set of $n \times n$ matrices with entries in $\mathbb{R}$. Then matrix addition $+$, is a commutative, associative binary operation on $M_n(\mathbb{R})$. It has an identity element, the zero matrix 0. Matrix multiplication $\cdot$ is also an associative binary operation on $M_n(\mathbb{R})$, however it is not commutative if $n > 1$. Despite being non-commutative, it does have an identity, the identity matrix I_n. (All this is again true if we take matrix entries in other number systems such as $\mathbb{Z}$, $\mathbb{Q}$ or $\mathbb{C}$. However if we take entries in $\mathbb{N}$, then we still have addition and multiplication, but there is no additive identity.)

Note that while we can also define matrix multiplication for pairs of non-square matrices of appropriate sizes (e.g., multiply a 2×3 matrix with a 3×2 matrix), this does not yield a binary operation because if we try to include a non-square matrix A in a set S,

then $A \cdot A$ will not be defined. On the other hand, matrix addition does still give a binary operation on the set $M_{m,n}(\mathbb{R})$ of $m \times n$ real matrices for any $m, n \in \mathbb{N}$.

Example 1.2.6. Let $\mathcal{P}(x;\mathbb{Z})$ be the space of polynomials in a single variable x with coefficients in $\mathbb{Z}$. Then polynomial addition and polynomial multiplication are commutative, associative binary operations with identity elements (namely the constant polynomials 0 and 1). The same is true for the space of polynomials over $\mathbb{Q}$ or $\mathbb{R}$ or $\mathbb{C}$.

There are loads of other operations on these spaces of polynomials as well. For instance, composition is a binary operation: $(f \circ g)(x) = f(g(x))$. This is not commutative but it is associative with identity (see exercise below). Another example of a binary operation is $f * g = f\frac{d}{dx}g + g\frac{d}{dx}f$, which comes up in the product rule for differentiation. (This example is commutative.)

These examples generalize in various ways. Addition and multiplication generalize to binary operations on polynomials with several variables, though composition does not. Addition, multiplication and composition generalize to binary operations on functions from $\mathbb{R} \to \mathbb{R}$. The example involving differentiation also gives a binary operation on smooth (infinitely differentiable) functions from $\mathbb{R}$ to $\mathbb{R}$.

Exercise 1.2.7. Consider the composition operation $\circ$ on $\mathcal{P}(x;\mathbb{Z})$. Show is it non-commutative, but that is has an identity element. What is the identity element?

1.2.2 Ring and field axioms

Definition 1.2.2. *Let R be a set with two binary operations, which we call addition $+ : R \times R \to R$ and multiplication $\cdot : R \times R \to R$. We say R is a* **(commutative) ring**[10] *if the following five properties (or axioms) hold:*

(1) $+$ *and* $\cdot$ *are associative;*

(2) $+$ *and* $\cdot$ *are commutative;*

(3) $+$ *and* $\cdot$ *satisfy the following (left)* distributive law*:*

$$a(b+c) = ab + ac,^{11} \quad \text{for all } a, b, c \in R; \text{ and} \tag{1.2.4}$$

(4) $+$ *and* $\cdot$ *have identity elements, denoted* 0 *and* 1 *respectively;*

(5) for each $a \in R$, there exists an element $-a \in R$, called the **additive inverse** *of a, such that $a + (-a) = 0$;*

If in addition, R has more than 1 element[12] *and R satisfies the following property:*

[10] If one wants to be more technical, one says the triple $(R, +, \cdot)$ is a ring, and R is the underlying set.

[11] Just like for multiplication of ordinary numbers, we often omit the $\cdot$ when writing ring multiplication, e.g., ab means $a \cdot b$.

[12] This is just a convention for technical reasons similar to the convention that 1 is not prime. If we allowed a field to have only 1 element, then many theorems about fields would need to exclude this degenerate case.

(6) *for each* nonzero $a \in R$, *there exists an element* $a^{-1} \in R$, *called the* **multiplicative inverse** *of* a, *such that* $a \cdot a^{-1} = 1$;

we say R *is a* **field**.

Warning: 0 and 1 are just notation for the additive and multiplicative identities. In general they are not the same as the usual integers 0 and 1.

Rings and fields should be properly treated in a course on algebra, and we will not go through all the formalities of checking all the axioms hold for our examples. What is important for us is to get a feel for what sort of things are rings and fields. Intuitively, being a ring means the following. Essentially, a ring is a number system where you have three operations: $+$, $-$ and $\cdot$ that satisfy expected properties.

Specifically, ring axioms (1)–(3) tell us that $+$ and $\cdot$ satisfy all the nice properties you're used to for addition and multiplication (see below). Axiom (4) essentially says your number system contains 0 and 1 (so $\mathbb{N}$ cannot be a ring). Recall from Exercise 1.2.4 that 0 and 1 are uniquely determined by the identity element property Eq. (1.2.3). (Caution: axiom (4) does not say that 0 and 1 are distinct elements of R—see below.) Axiom (5) says that "negatives" exist in the ring (so $\mathbb{Z}_{\geq 0}$ cannot be a ring). Therefore we can subtract in the ring according to $a - b = a + (-b)$, and negation behaves as expected.

Similarly, the essential idea of what a field is is a number system where you have four operations: $+$, $-$, $\cdot$ and $/$ that satisfy all the usual properties, where we define $a/b = ab^{-1}$ when $b \neq 0$.

To be more concrete about what I mean by the axioms implying $+, -$ and $\cdot$ have the properties you would expect, consider the following proposition, listing many but not all the properties we're familiar with from usual arithmetic.

Proposition 1.2.3. *Let* R *be a ring and* $a, b, c \in R$. *Then the following properties hold:*

(1) *[cancellation]* $a + c = b + c \implies a = b$;

(2) *[uniqueness of additive identity]* $a + b = 0 \implies b = -a$;

(3) *[uniqueness of multiplicative identity]* $ab = 1 \implies a^{-1}$ *exists and* $b = a^{-1}$;

(4) *[double negation]* $-(-a) = a$;

(5) *[double inversion] if* a^{-1} *exists, then* $(a^{-1})^{-1}$ *exists and it equals* a;

(6) *[right distributive law]* $(a + b)c = a \cdot c + b \cdot c$;

(7) *[multiplication by 0]* $0 \cdot a = 0$;

(8) *[commutativity of negation]* $(-a)b = (-b)a$;

(9) *[cancellation of negations]* $(-a)(-b) = ab$;

(10) *[distribution of subtraction]* $a(b - c) = ab - ac$; *and*

number system	ring?	field?
$\mathbb{N}$	no	no
$\mathbb{Z}_{\geq 0}$	no	no
$\mathbb{Z}$	yes	no
$\mathbb{Q}$	yes	yes
$\mathbb{R}$	yes	yes
$\mathbb{C}$	yes	yes

Table 1.2: Ring/field classification of standard number systems

(11) *[distribution of negation]* $a-(b+c)=a-b-c$.

Proof. I'll just exhibit proofs for properties (1) and (7), and let you complete the rest if you desire.

Property (1) follows as

$$\begin{aligned} & a+c=b+c \\ \text{(Axiom 5)} \quad \implies & (a+c)+(-c)=(b+c)+(-c) \\ \text{(Axiom 1)} \quad \implies & a+(c+(-c))=b+(c+(-c)) \\ \text{(Axiom 5)} \quad \implies & a+0=b+0 \\ \text{(Axiom 4)} \quad \implies & a=b. \end{aligned}$$

The proof for property (7) goes along the lines of an argument you may have seen in linear algebra, and uses the fact that $0+0=0$, which follows from the definition of an additive identity.

$$\begin{aligned} 0\cdot a &= (0+0)\cdot a \\ \text{(property (6))} \quad &= 0\cdot a+0\cdot a. \end{aligned}$$

This implies $0\cdot a+0\cdot a=0\cdot a+0$ (by definition of additive identity), so $0\cdot a=0$ by property (1). □

Exercise 1.2.8. Prove properties (4), (6) and (9) of the above proposition.

Now let's look at some examples.

Example 1.2.7. $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ are all rings.[13] All but $\mathbb{Z}$ are fields (only ± 1 have multiplicative inverses in $\mathbb{Z}$), as any nonzero element a of $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$ has a multiplicative inverse (i.e., a reciprocal, $\frac{1}{a}$) which is again in the ring. We tabulate this information, along with the earlier comments about $\mathbb{N}$ and $\mathbb{Z}_{\geq 0}$ not being rings (and hence not fields either), in Table 1.2.

[13] I am taking all properties listed in the ring axioms for granted for these standard number systems. One can verify these formally starting from Peano's axioms and the constructions of the other number systems from $\mathbb{N}$, but it's tedious and I think not really enlightening (with the possible exception of how things go for $\mathbb{R}$). Anyway, I don't want this course to be that kind of course.

Example 1.2.8. $\mathbb{Q}^\times$, $\mathbb{R}^\times$ and $\mathbb{C}^\times$ (cf. Example 1.2.4) are not rings because they do not possess 0.

Example 1.2.9. Let $R = \{0\}$. There is only a single binary operation on a set with one element: $0 * 0 = 0$. We let both $+$ and $\cdot$ denote this operation. Then R is a ring (you can check the axioms if you like), called the **zero ring**. In this ring, $1 = 0$. (Note 1 is just the notation for the multiplicative identity—since $0 \cdot 0 = 0$ and 0 is the only element in R, this means 0 is the multiplicative identity—cf. warning above.) Note that R satisfies all 6 field axioms, but we defined fields to have more than 1 element, so this is not a field.

This funny situation where $0 = 1$ can only happen in the zero ring. In particular, in any field $0 \neq 1$.

Exercise 1.2.9. Let R be a ring with more than one element. Prove that $1 \neq 0$. (*Hint:* Try contradiction and use property (7) of Proposition 1.2.3.)

Example 1.2.10. The set of integral polynomials $\mathcal{P}(x; \mathbb{Z})$ is a ring: you can add, subtract and multiply polynomials, and these operations satisfy the usual properties. This is also true for polynomials in several variables, and one can take coefficients in $\mathbb{Q}$ or $\mathbb{R}$ or $\mathbb{C}$ as well. However, even if one takes coefficients in a field, say $\mathbb{Q}$, the ring of polynomials is not a field: e.g., $\frac{1}{x+1}$ is not a polynomial.

Similarly, the functions from $\mathbb{R}$ to $\mathbb{R}$ form a ring, but not a field, as any function $f(x)$ such that $f(a) = 0$ for some $a \in \mathbb{R}$ cannot have a multiplicative inverse. (If $f \cdot g = 1$, then $1 = f(a)g(a) = 0 \cdot g(a)$, which is impossible.)

Example 1.2.11. For $n > 1$, the space $M_n(\mathbb{R})$ of $n \times n$ real matrices is not a commutative ring, because matrix multiplication is not commutative. However, it is what is known as a *non-commutative ring*.[14] We'll discuss non-commutative rings a little later, but just say now that matrix rings are the prototypical example of non-commutative rings.

Example 1.2.12. Let $V = \mathbb{R}^n$. Then V is an n-dimensional real vector space, which has 2 operations: addition and *scalar* multiplication. Addition is a binary operation on V, but scalar multiplication is not (at least if $n > 1$)—rather scalar multiplication is a map $\mathbb{R} \times V \to V$. For vector spaces, there is no natural way to multiply two vectors and get another vector, so V is not naturally a ring. (In the special case of $\mathbb{R}^3$, we could try to define multiplication of vectors with the cross product, but this still will not make a ring—e.g., there is no vector e such that $e \times v = v \times e = v$ for all $v \in \mathbb{R}^3$.)

That said, one can make V into a ring by defining a suitable multiplication. In fact there are several ways to do this. The most obvious one is by component-wise multiplication, i.e.,

$$(u_1, u_2, \cdots, u_n) \cdot (v_1, v_2, \cdots, v_n) = (u_1 v_1, u_2 v_2, \cdots, u_n v_n)$$

[14] Outside of this class, the word "ring" usually means a commutative or a non-commutative ring, so outside of this class people including me will simply say that $M_n(\mathbb{R})$ is a ring.

makes V into a commutative ring. (What are the additive and multiplicative identities? Is it a field?) However, it is not a particularly interesting one.

When $n = 2$, we can identify $V = \mathbb{R}^2$ with $\mathbb{C}$ via $(x, y) = x + iy$ and use this to define a multiplication on $\mathbb{R}^2$, which is different from the component-wise definition. (What is $(x, y) \cdot (u, v)$ in this situation?) Also, in the special case $n = m^2 \geq 4$ for some $m \in \mathbb{N}$, then we can identify V with $M_m(\mathbb{R})$ making V into a non-commutative ring. It turns out that many of the rings arising in number theory (e.g., the quadratic rings in Section 1.5 or the quaternions mentioned in Section 1.7) can be viewed as putting interesting multiplication rules on vector spaces, though we will not emphasize this.

From the latter examples, we see that the notion of a ring is actually more general than what you think the term "number system" should mean. (Who thinks of polynomials as number systems?) So perhaps it's better to think of rings and fields as generalizations of number systems where one can do arithmetic. However, for the most part the rings and fields we will be considering in this class, even besides the standard number systems, really are some type of number systems. In the next sections, we will introduce some of these other number systems such as the integers mod n and the Gaussian integers, which will feature as *dramatis personae* in this course.

Before we move onto these other number systems, let us give a useful method to determine if certain objects are rings or fields. One can of course check the definition by checking all the axioms, but that is rather tedious.

Definition 1.2.4. *Let $S \subset R$. If S and R are rings (with respect to the same operations $+$ and $\cdot$), we say S is a* **subring** *of R. Similarly, if S and R are fields (with respect to the same operations $+$ and $\cdot$), we say S is a* **subfield** *of R.*

Example 1.2.13. $\mathbb{Z}$ is a subring of $\mathbb{Q}$, which is a subfield of $\mathbb{R}$, which is a subfield of $\mathbb{C}$.

Example 1.2.14. We can view $\mathbb{Z}$ as a subring of $\mathcal{P}(x; \mathbb{Z})$ (by identifying an integer a with the constant polynomial $f(x) = a$). On the other hand, $\mathbb{Q}$ (or $\mathbb{R}$ or $\mathbb{C}$) is not a subring of $\mathcal{P}(x; \mathbb{Z})$, and vice versa.

Definition 1.2.5. *Let R be a ring, and $S \subset R$. Let $*$ be one of the operations $+, -, \cdot$. We say S* **closed under** *$*$ if $a * b \in S$ for all $a, b \in S$. If R is a field, we say S is closed under division by nonzero elements if $a/b \in S$ for all $a, b \in S$ with $b \neq 0$.*

Example 1.2.15. Let $R = \mathbb{Z}$. Then $S = \mathbb{N}$ is closed under $+$ and $\cdot$, but not under $-$.

Example 1.2.16. Let $R = \mathbb{Z}$, and $S = 2\mathbb{Z}$, the set of even integers. Then S is closed under $+$ (the sum of two even numbers is even), $-$ (the difference of two even numbers is even) and $\cdot$ (the product of two even numbers is even). On the other hand, we see the set of odd integers is closed under $\cdot$, but not under $+$ or $-$.

Lemma 1.2.6. *Let R be a ring and $S \subset R$. Then S is a subring of R if and only if $1 \in S$ and S is closed under addition, subtraction and multiplication. Similarly if R is a field, S is a subfield of R if and only if S contains a nonzero element and is closed under addition, subtraction, multiplication and division by nonzero elements.*

The proof is similar to the test for subspaces you should have seen in Linear Algebra, and I will leave it as an exercise. The idea is that it's tedious to check that operations in something you want to be a ring are the commutative, associative and distributive (ring axioms (1)–(3)), but these come for free for S if we already know them for R. Then one checks the closure properties stated in the lemma imply (in fact, are equivalent to) the remaining ring axioms. Note by Example 1.2.16, we also need the condition $1 \in S$ to make the lemma true for subrings. (One doesn't need this for the field part, because closure under division by nonzero elements already implies $1 = a/a \in S$ for any $a \in S$, i.e. $1 \in S$ provided S has a non-zero element.)

The usefulness is that if we want to show S is a ring or a field, and we know already it is contained in another ring or field R, it suffices to check these closure properties. Here's a simple illustration:

Example 1.2.17. Let $\mathcal{C}(\mathbb{R};\mathbb{R})$ denote the space of continuous functions from $\mathbb{R}$ to $\mathbb{R}$. We've already stated (Example 1.2.10) that the set $\mathcal{F}(\mathbb{R},\mathbb{R})$ of all functions from $\mathbb{R}$ to $\mathbb{R}$ is a ring (and a proper proof is not too hard). Since $\mathcal{C}(\mathbb{R};\mathbb{R}) \subset \mathcal{F}(\mathbb{R},\mathbb{R})$, to check $\mathcal{C}(\mathbb{R};\mathbb{R})$ is a ring, by the above lemma, it suffices to check $1 \in \mathcal{C}(\mathbb{R};\mathbb{R})$ (it is as all constant functions are continuous), and $\mathcal{C}(\mathbb{R};\mathbb{R})$ is closed under $+$, $-$ and $\cdot$. Here one uses the theorems that the sum and product of continuous functions are continuous. This gives closure under $+$ and $\cdot$. Also, since -1 is continuous, for $f, g \in \mathcal{C}(\mathbb{R};\mathbb{R})$, $-g$ is continuous so $f - g = f + (-g) \in \mathcal{C}(\mathbb{R};\mathbb{R})$. Hence we have closure under $-$, and $\mathcal{C}(\mathbb{R};\mathbb{R})$ is a subring of $\mathcal{F}(\mathbb{R},\mathbb{R})$; in particular, it's a ring.

Exercise 1.2.10. Let

$$\mathbb{Z}[\tfrac{1}{2}] = \left\{ \frac{a}{b} \in \mathbb{Q} : b = 2^n \text{ for some } n \in \mathbb{Z}_{\geq 0} \right\}.$$

Show that $\mathbb{Z}[\frac{1}{2}]$ is a ring by showing it is a subring of $\mathbb{Q}$.

Exercise 1.2.11. Prove Lemma 1.2.6.

1.3 Integers mod n

The most basic and important number systems in number theory after the standard number systems are the integers mod n, for $n \in \mathbb{N}$, denoted $\mathbb{Z}/n\mathbb{Z}$, or sometimes for simplicity $\mathbb{Z}/n$.[15]

[15] Some authors use $\mathbb{Z}_n$ instead of $\mathbb{Z}/n\mathbb{Z}$, but when $n = p$ is prime, this contradicts with standard notation $\mathbb{Z}_p$ for the *p-adic integers* (see Section 1.7), so number theorists typically avoid that notation. Actually, $\mathbb{Z}/n$ is not really used in formal writing and I don't plan to use it in these typed notes at all, but it's cumbersome to write $\mathbb{Z}/n\mathbb{Z}$ by hand all the time, so I may sometimes write $\mathbb{Z}/n$ on the board as shorthand.

These systems are used for what is known as *modular arithmetic*, which probably you have seen before, say in Discrete Math. Modular arithmetic turns out to be supremely useful in number theory, and is also quite useful in computer science. We'll explore this in Chapter 3. For now, we'll just explain the number systems $\mathbb{Z}/n\mathbb{Z}$.

For integers a, b, we write $a \mid b$ for a divides b, i.e., b is an (integer) multiple of a, i.e., $b = ka$ for some $k \in \mathbb{Z}$. In particular, every integer divides 0, and ± 1 divides every integer.

Definition 1.3.1. *Let $a, b, n \in \mathbb{Z}$. We say a and b are* **congruent** mod n *(or* **congruent modulo** n *or* **equivalent** mod n*), and write $a \equiv b \bmod n$, if $n \mid (b-a)$, i.e., if $b - a$ is a multiple of n.*

You should've learned in Discrete Math that congruence mod n is an equivalence relation. In case you didn't or you forgot, prove it:

Exercise 1.3.1. Let $n \in \mathbb{Z}$. Show that congruence mod n is an equivalence relation, i.e., we have:

(i) $a \equiv a \bmod n$ for all $a \in \mathbb{Z}$;

(ii) $a \equiv b \bmod n$ implies $b \equiv a \bmod n$ for all $a, b \in \mathbb{Z}$; and

(iii) $a \equiv b \bmod n$ and $b \equiv c \bmod n$ implies $a \equiv c \bmod n$.

Equivalence relations partition sets into equivalence classes.

Definition 1.3.2. *Let $a, n \in \mathbb{Z}$. The* **equivalence** *(or* **congruence***)* **class** *of a mod n is*

$$n\mathbb{Z} + a = a + n\mathbb{Z} = \{a + kn : k \in \mathbb{Z}\} = \{\dots, a - 2n, a - n, a, a + n, a + 2n, \dots\}.$$

(Whether we write $n\mathbb{Z} + a$ or $a + n\mathbb{Z}$ is simply a matter of preference.)

Note that $n\mathbb{Z} + a = \{b \in \mathbb{Z} : a \equiv b \bmod n\}$, i.e., this is the set of all elements which are equivalent to a mod n, which is the usual way to define equivalence classes in general. In particular, $n\mathbb{Z} + 0$, the equivalence class of 0 mod n, simply means the multiples of n. We often denote $n\mathbb{Z} + 0$ simply by $n\mathbb{Z}$.

Exercise 1.3.2. Let $n \in \mathbb{Z}$. Show that $n\mathbb{Z} + a = \{b \in \mathbb{Z} : a \equiv b \bmod n\}$.

The equivalence classes are infinite except in the special case that $n = 0$, when the have size 1: $0\mathbb{Z} + a = \{a\}$. (Below, we often assume $n > 0$.) Two equivalence classes $n\mathbb{Z} + a$ and $n\mathbb{Z} + b$ are the same if and only if $a \equiv b \bmod n$. The equivalence classes partition $\mathbb{Z}$ as follows.

Proposition 1.3.3. *Let $n \in \mathbb{N}$. Then there are n distinct equivalence classes mod n:*

$$n\mathbb{Z},\ n\mathbb{Z} + 1,\ n\mathbb{Z} + 2,\ \dots,\ n\mathbb{Z} + (n-1),$$

and

$$\mathbb{Z} = n\mathbb{Z} \sqcup (n\mathbb{Z} + 1) \sqcup \cdots \sqcup (n\mathbb{Z} + (n-1)).$$

Here $\sqcup$ denotes the disjoint union, which is the same as the usual set union $\cup$, except it carries the additional connotation that all of the sets being unioned are all disjoint (no two have any elements in common).

Proof. By general properties of equivalence relations, no $a \in \mathbb{Z}$ can lie in two distinct equivalence classes. (If this is not familiar to you, convince yourself it's true for equivalence mod n.) Since every $a \in \mathbb{Z}$ lies in some equivalence class, $\mathbb{Z}$ must be the disjoint union of the equivalence classes mod n. Hence it suffices to prove that the above list is a complete set of distinct equivalence classes.

We first want to show that every equivalence class is one of the n equivalence classes given above, i.e., of the form $n\mathbb{Z} + a$ for some $0 \leq a < n$. Consider an arbitrary equivalence class $C = n\mathbb{Z} + a$, for $a \in \mathbb{Z}$. If $0 \leq a < n$, then we are done.

Suppose $a \geq n$. Then we can also write $C = n\mathbb{Z} + a'$ where $a' = a - n$. Note $0 \leq a' < a$. If $a' < n$, we are done. If not we can repeat this procedure and write $C = n\mathbb{Z} + a''$ where $a'' = a' - n$, and $0 \leq a'' < a' < a$. Continuing in this manner, we eventually get some representation $C = n\mathbb{Z} + a^{(k)}$ where $0 \leq a^{(k)} < n$ by the descent principle. (Concretely $a^{(k)}$ is the remainder upon division of a by n. The above argument actually proves that one gets a remainder upon division, which is why we gave it.)

The case where $a < 0$ follows by a similar argument, and we can conclude that any equivalence class mod n is of the form $C = n\mathbb{Z} + a$ with $0 \leq a < n$. However, we are not quite done. We haven't proven that all of these n equivalence classes are actually distinct.

To do this, suppose $n\mathbb{Z} + a = n\mathbb{Z} + b$ with $0 \leq a, b < n$. Then $b \in n\mathbb{Z} + a$, i.e., $a \equiv b \bmod n$, i.e., $n \mid (b - a)$. We may assume $b \geq a$ (otherwise, interchange a and b). Then $0 \leq b - a < n$ and $b - a$ is a multiple of n. The only possibility is $b - a = 0$, i.e., $a = b$ as desired. □

Example 1.3.1. Suppose $n = 2$. Then the equivalence classes of $\mathbb{Z}$ mod 2 are

$$2\mathbb{Z} = \{\ldots, -4, -2, 0, 2, 4, 6, \ldots\}$$
$$2\mathbb{Z} + 1 = \{\ldots, -3, -1, 1, 3, 5, 7, \ldots\}.$$

Example 1.3.2. Suppose $n = 3$. Then the equivalence classes of $\mathbb{Z}$ mod 3 are

$$3\mathbb{Z} = \{\ldots, -6, -3, 0, 3, 6, 9, \ldots\}$$
$$3\mathbb{Z} + 1 = \{\ldots, -5, -2, 1, 4, 7, 10, \ldots\}$$
$$3\mathbb{Z} + 2 = \{\ldots, -4, -1, 2, 5, 8, 11, \ldots\}.$$

Definition 1.3.4. *Let $n \in \mathbb{Z}$. Denote the set of equivalence classes mod n by $\mathbb{Z}/n\mathbb{Z}$, which we call* **the integers mod** n. *We define binary operations of addition and multiplication on $\mathbb{Z}/n\mathbb{Z}$ as follows. The sum of two equivalence classes is given by*

$$(n\mathbb{Z} + a) + (n\mathbb{Z} + b) = n\mathbb{Z} + (a + b).$$

The product is given by

$$(n\mathbb{Z} + a)(n\mathbb{Z} + b) = n\mathbb{Z} + ab.$$

The idea of modular arithmetic is very simple. A naive definition of a modulo n is the remainder upon division by n, which is the unique number $0 \leq b < n$ such that $a \equiv b \bmod n$. Then one can think of the "integers modulo n" as the possible remainders upon division $0, 1, \ldots, n-1$. Often we will want to do arithmetic mod n, e.g., figuring out what time it will be 9 hours after 5 o'clock on a 12-hour clock means computing $9+5$ modulo 12. In this example, we would like to say that

$$(9 \bmod 12) + (5 \bmod 12) = 2 \quad (= 2 \bmod 12).$$

However this is technically incorrect as, with our naive definition,

$$(9 \bmod 12) + (5 \bmod 12) = 9 + 5 = 14 \neq 2.$$

In other words, usual integer addition doesn't make sense on the naive version of integers mod n (it's not a binary operation). So one approach is to write equations as mod n equivalence relations, e.g.,

$$9 + 5 \equiv 14 \equiv 2 \bmod 12.$$

In fact, this is what we are doing by defining addition and multiplication on $\mathbb{Z}/n\mathbb{Z}$ (equivalence classes)—we are just using more sophisticated language because this will be useful for understanding modular arithmetic theoretically. Namely, in terms of how we defined addition on $\mathbb{Z}/n\mathbb{Z}$, we can translate the above equation as the following addition statement on $\mathbb{Z}/12\mathbb{Z}$:

$$(12\mathbb{Z} + 9) + (12\mathbb{Z} + 5) = 12\mathbb{Z} + 14 = 12\mathbb{Z} + 2.$$

The point of this more sophisticated language is that it will be useful for a theoretical understanding of modular arithmetic. That said, when doing actual calculations, it is cumbersome to write elements of $\mathbb{Z}/n\mathbb{Z}$ as $n\mathbb{Z}+a$. So we will often refer to elements of $\mathbb{Z}/n\mathbb{Z}$ (e.g., a class $n\mathbb{Z} + a$) by a representative of the class (e.g., a). For instance, we may refer to the class $12\mathbb{Z} + 9$ as the element 9 in $\mathbb{Z}/12\mathbb{Z}$, or we might call it the element -3 in $\mathbb{Z}/12\mathbb{Z}$ depending on which is convenient, with it being understood that we mean the class of integers containing that number. Nevertheless, to avoid confusion we will not write *explicit* equations in this form, e.g., we will not write $9+5 = 2$ (in $\mathbb{Z}/12\mathbb{Z}$), but rather in congruence notation: $9 + 5 \equiv 2 \bmod 12$.

Thinking about clocks as an example of modular arithmetic suggests a way to visualize $\mathbb{Z}/n\mathbb{Z}$ in general, as n points on a circle, e.g., for $n = 6$ see Fig. 1.3.1. This picture is very suggestive of the "wrap-around" structure of addition mod n. For instance, adding one just moves you one position clockwise around the circle. You can also use this visualize multiplication mod n. E.g., if we want to compute $5 \cdot 4 \bmod 6$, we can think of taking 5 steps (clockwise) around the circle with step size 4 (starting from 0). I.e., our first step puts at 4, the next step at $4 + 4 \equiv 2 \bmod 6$ and so on until we end up back at 2.

The following result tells us that modular arithmetic has nice properties.

Theorem 1.3.5. *Let $n \in \mathbb{Z}$. Then $\mathbb{Z}/n\mathbb{Z}$ is a ring.*

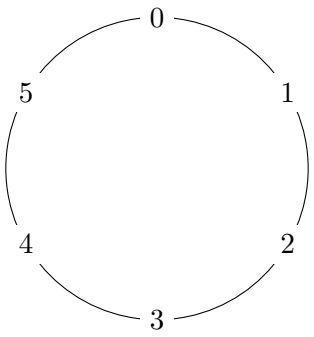


Figure 1.3.1: Visualizing $\mathbb{Z}/6\mathbb{Z}$

Observe that this picture of $\mathbb{Z}/n\mathbb{Z}$ looks like a *ring* (in the non-mathematical sense), so the mathematical usage of the word ring now may make some sense.[16]

Proof. It is not to hard to check the ring axioms, e.g., one can show that ring axioms (1)–(3) hold in $\mathbb{Z}/n\mathbb{Z}$ because they do in $\mathbb{Z}$. (However one cannot directly use Lemma 1.2.6 since $\mathbb{Z}/n\mathbb{Z}$ is not a subset of $\mathbb{Z}$.) The main thing to check is that the above definitions of $+$ and $\cdot$ on $\mathbb{Z}/n\mathbb{Z}$ are actually well defined. Namely, we have described how to add and multiply equivalence classes, but the description a priori depends on a choice of description of these equivalence classes. We need to show it in fact does not.

Consider two equivalence classes C and D in $\mathbb{Z}/n\mathbb{Z}$, any two representations of each equivalence class, say:

$$C = n\mathbb{Z} + a = n\mathbb{Z} + a'$$
$$D = n\mathbb{Z} + b = n\mathbb{Z} + b',$$

for some $a, a', b, b' \in \mathbb{Z}$. The above definition of addition of equivalence classes tells us both

$$C + D = n\mathbb{Z} + (a + b) \quad \text{and} \quad C + D = n\mathbb{Z} + (a' + b').$$

Showing addition is well defined, i.e., does not depend upon our representation of C and D, means that showing these two equivalence classes are the same, i.e., that $a+b \equiv a'+b' \bmod n$.

Since $a \equiv a' \bmod n$ and $b \equiv b' \bmod n$, we can write $a' = jn + a$ and $b' = kn + b$ for some $j, k \in \mathbb{Z}$. Hence $a' + b' = jn + kn + a + b = (j + k)n + (a + b)$. Therefore $a + b \equiv a' + b' \bmod n$, and we have shown addition in $\mathbb{Z}/n\mathbb{Z}$ is well defined. The case of multiplication is an exercise.

Now the ring axioms follow easily from those for $\mathbb{Z}$. Convince yourself of axioms (1)–(3) now. Axioms (4) and (5) are an exercise below. □

[16] It appears the mathematical term *ring* was introduced by Hilbert in the 1890s, possibly with a similar but more general sort of notion in mind. We remark the term for field in German is *Körper* (the modern definition of field was essentially introduced by Dedekind in 1858 in German) French is *corps*, both of which mean "body" or "corpus." Anyway, I would not try to read too much into this terminology.

Exercise 1.3.3. Show multiplication in $\mathbb{Z}/n\mathbb{Z}$ is well defined.

Exercise 1.3.4. Let $n \in \mathbb{N}$. Show that $\mathbb{Z}/n\mathbb{Z}$ satisfies ring axioms (4) and (5).

Example 1.3.3. Addition and multiplication tables for $\mathbb{Z}/2\mathbb{Z}$ are given by

$+$	0	1
0	0	1
1	1	0

and

$\cdot$	0	1
0	0	0
1	0	1

Note that there is only one nonzero element of $\mathbb{Z}/2\mathbb{Z}$, namely 1, which has multiplicative inverse 1, so $\mathbb{Z}/2\mathbb{Z}$ is also a field.

Example 1.3.4. Addition and multiplication tables for $\mathbb{Z}/3\mathbb{Z}$ are given by

$+$	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

and

$\cdot$	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

The nonzero elements are 1 and 2 (which is also -1) in $\mathbb{Z}/3\mathbb{Z}$. Their multiplicative inverses are 1 and 2. Hence $\mathbb{Z}/3\mathbb{Z}$ is also a field.

However, $\mathbb{Z}/n\mathbb{Z}$ is not always a field. We'll consider a few more questions now, and settle the question of when $\mathbb{Z}/n\mathbb{Z}$ is a field completely in Chapter 3.

Example 1.3.5. $\mathbb{Z}/4\mathbb{Z}$ is not a field. Namely $2 \in \mathbb{Z}/4\mathbb{Z}$ does not have a multiplicative inverse. We can prove this by contradiction: Suppose it does, i.e., $2a \equiv 1 \bmod 4$ for some $a \in \mathbb{Z}$. Then $4|(2a-1)$, but $2a-1$ is odd, a contradiction.

Exercise 1.3.5. Write down addition and multiplication tables for $\mathbb{Z}/5\mathbb{Z}$. Which elements have a multiplicative inverse? Is $\mathbb{Z}/5\mathbb{Z}$ a field?

Exercise 1.3.6. Write down addition and multiplication tables for $\mathbb{Z}/6\mathbb{Z}$. Which elements have a multiplicative inverse? Is $\mathbb{Z}/6\mathbb{Z}$ a field?

The fact that $\mathbb{Z}/n\mathbb{Z}$ is a ring gives us a simple way to do computations mod n.

For instance, if we want to add several numbers mod n, e.g., the primes between 10 and 20 mod 5:

$$11 + 13 + 17 + 19 \bmod 5,$$

it suffices to add their equivalence classes together mod 5, i.e.,

$$11 + 13 + 17 + 19 \equiv 1 + 3 + 2 + 4 \equiv 1 + (-2) + 2 + (-1) \equiv 0 \bmod 5.$$

This gives a computationally easier way to check that $11 + 13 + 17 + 19$ is divisible by 5 than first computing the sum as an integer (which is 60) and checking divisibility.

Similarly, if we wanted to compute 5^{10} mod 3, it suffices to multiply the equivalence class of 5 mod 3 to itself 10 times, and we see

$$5^{10} \equiv (-1)^{10} \equiv 1 \bmod 3.$$

It's a common trick in modular arithmetic to sometimes use representatives which are negative.

In general if we want to do a bunch of arithmetic operations and take the result mod n, we can work mod n at all intermediate steps, which often makes life much easier.[17]

Warning: don't take the exponent mod n in this process, only numbers that you are adding, subtracting, multiplying or dividing — e.g., in above example of 5^{10} mod 3, the exponent is 1 mod 3 but $5^{10} \equiv 1 \not\equiv 5^1 \equiv 2 \bmod 3$. In Chapter 3 we'll learn how to reduce exponents in modular arithmetic.

Example 1.3.6. What are the last two digits of $97 \cdot 98 \cdot 99$?

To find the last two digits of a number a, we just need to compute a mod 100. Note

$$97 \cdot 98 \cdot 99 \equiv (-3)(-2)(-1) \equiv -6 \equiv 94 \bmod 100,$$

so the last two digits are 94.

Exercise 1.3.7. Compute by hand 9^9 mod 7.

[17] The more sophisticated way to view this statement, once you know a little ring theory, is to say that the reduction mod n map from $\mathbb{Z}$ to $\mathbb{Z}/n\mathbb{Z}$ is a *ring homomorphism*.

Exercise 1.3.8. Let k be the product of all odd numbers from 1 to 99 (inclusive). What is k mod 4?

Exercise 1.3.9. Let k be the sum of all odd numbers from 1 to 99 (inclusive). What is the last digit of k?

While we defined the integers mod n for negative n, this doesn't actually give you anything different than integers mod positive numbers:

Exercise 1.3.10. Let $n \in \mathbb{N}$. Show that the integers mod n are the same as the integers mod $-n$. Namely, for any $a \in \mathbb{Z}$ we have $a + n\mathbb{Z} = a + (-n)\mathbb{Z}$.

1.4 Lapine numbers

Not all creatures have the same conception of numbers as we sophisticated humans. Rabbits essentially have five numbers: one, two, three, four, and hrair.[18] Hrair means many, and it's used for any counting number greater than 4. Let's briefly look at how this gives us an alternative number system. While this may seem like nonsense, I think it is actually instructive, for a couple of reasons. First, humans actually aren't all that good at intuitively understanding really large numbers, and this gives us a toy model for framing such ideas. Second, just like you must know night to understand day, you often better understand something in mathematics by knowing what it is not (this is why I gave you examples of non-rings as well as rings after defining rings). I think by looking at this lapine number system will help us appreciate some features of $\mathbb{N}$ that you may take for granted. In addition, understanding a technical issue with this system may illuminate the definition of modular arithmetic.

Let $\mathbb{L} = \{1, 2, 3, 4, \text{hrair}\}$. Addition and multiplication are binary operations on $\mathbb{L}$ given by the following tables:

$+$	1	2	3	4	hrair
1	2	3	4	hrair	hrair
2	3	4	hrair	hrair	hrair
3	4	hrair	hrair	hrair	hrair
4	hrair	hrair	hrair	hrair	hrair
hrair	hrair	hrair	hrair	hrair	hrair

and

$\cdot$	1	2	3	4	hrair
1	1	2	3	4	hrair
2	2	4	hrair	hrair	hrair
3	3	hrair	hrair	hrair	hrair
4	4	hrair	hrair	hrair	hrair
hrair	hrair	hrair	hrair	hrair	hrair

[18] *Watership Down*, by Richard Adams. Or for a possibly less sexist reference: *Tales from Watership Down*.

It is clear from the tables that $+$ and $\cdot$ are commutative.

Exercise 1.4.1. Show that addition and multiplication on $\mathbb{L}$ are associative.

What about subtraction and division? For subtraction, we don't get a binary operation, but we can make still make a table:

$-$	1	2	3	4	hrair
1	0	$-$	$-$	$-$	$-$
2	1	0	$-$	$-$	$-$
3	2	1	0	$-$	$-$
4	3	2	1	0	$-$
hrair	$-$	$-$	$-$	$-$	$-$

Here an entry of $-$ means the operation $x-y$ is undefined (recall our convention for operation tables is $x - y$ is represented by the entry in row x and column y). I've used 0 as distinct from $-$, even though 0 technically isn't in $\mathbb{L}$, because surely rabbits know that if you have 3 cabbages, and you take away 3 cabbages, you have no cabbages left. (Technically, one can define an extended lapine number system $\mathbb{L}' = \{0, 1, 2, 3, 4, \text{hrair}\}$ and work with operations on them, but I don't want my operation tables to get any bigger.)

Exercise 1.4.2. Make an analogous table for division on $\mathbb{L}$.

Now if rabbits were clever enough to understand negative numbers,[19] one could similarly fill in some more of the undefined values in the subtraction table, such as $1 - 2 = -1$. However, the hrair row and the hrair column cannot be defined. To explain this in as complicated a way as possible, let's think about how $\mathbb{L}$ relates to $\mathbb{N}$. We can think of hrair as being the collection of all numbers bigger than 4:

$$\text{hrair} = \{5, 6, 7, \ldots\}.$$

To uniformly think of all lapine numbers as subsets of $\mathbb{N}$, we will think of the lapine numbers 1, 2, 3 and 4 as being the singleton sets $\{1\}$, $\{2\}$, $\{3\}$ and $\{4\}$. In this way, the elements of give a partition of $\mathbb{N}$ into 5 subsets, hence corresponds to an equivalence relation $\sim$ on $\mathbb{N}$: namely the relation that $a \sim b$ if $a = b$ or if $a, b \geq 5$. Then the elements of are simply the equivalence classes of $\mathbb{N}$.

Note that addition and multiplication make sense on these equivalence classes: for any $A, B \in \mathbb{L}$ (thinking of A, B as subsets of $\mathbb{N}$), we define $A+B$ (resp. $A \cdot B$) to be the equivalence class containing $a + b$ (resp. $a \cdot b$) for some $a \in A, b \in B$. For instance $\{1\} + \{3\} = \{4\}$ since $1 + 3 = 4$, $\{2\} + \{3\} = \text{hrair}$ because $2 + 3 \in \text{hrair}$ and $\text{hrair} + \text{hrair} = \text{hrair}$ because $a + b \in \text{hrair}$ for $a, b \in \text{hrair}$. The key point is that these operations are well defined because they do not depend upon the choice of $a \in A$ and $b \in B$. We needed the same property to define modular arithmetic—recall the proof of Theorem 1.3.5.

[19] Rabscuttle or Blackberry might be.

However, this is not true for subtraction (or division). Imagine trying to define subtraction in the same way: $A - B = C$ where C is the equivalence class containing $a - b$ for some $a \in A$, $b \in B$. Then what is hrair $-$ hrair? Here we take $A = B =$ hrair. Say $b = 5$. If we chose $a = 6$, we get hrair $-$ hrair $= 1$, but if we chose $a = 7$ we get hrair $-$ hrair $= 2$. In fact we could get hrair $-$ hrair is $1, 2, 3, 4$ or hrair, or even something not in $\mathbb{L}$ (if $a - b \leq 0$). This is similar to how $\infty - \infty$ is an indeterminate form in calculus, and we could think of hrair $-$ hrair as an indeterminate form in $\mathbb{L}$. (More similarly, we think of $\infty + \infty = \infty$ as making sense just like hrair $+$ hrair $=$ hrair makes sense.) On the other hand, with this definition of subtraction quantities like $\{3\} - \{2\} = \{1\}$ still make sense because when A and B are singleton sets there is only one choice for representatives $a \in A$ and $b \in B$. Put another way, the problem with the above definition is that we said C is *the* equivalence class containing $a - b$, which implies that C is uniquely determined, but it is not always. (There is no such issue with subtraction for modular arithmetic: think about it.)

Exercise 1.4.3. Show that hrair $- B$ is not well defined (an indeterminate form) for any $B \in \mathbb{L}$. For each $B \in \mathbb{L}$, explicitly determine the possibilities for hrair $- B$ according to the above definition.

I think this model of $\mathbb{L}$ is not so different from how we actually intuitively understand numbers. Yes, we can count way past 4 without much trouble, but we have trouble understanding the scale of large numbers—particularly if they're given different formats. Do you have any sense of how big $2^{100,000} - 10,000!$ is? You shouldn't, unless there's something seriously wrong with you. It's not even immediately clear if it's negative or positive. The answer is it's negative, and is approximately equal to $-10,000!$, since $10,000!$ has 35,660 digits and $2^{100,000}$ has a mere 30,103 digits. Even from this information, how do you intuitively understand the difference in scale between $10,000!$ and $2^{100,000}$? They're both ridiculously huge, and both have 30-some thousand digits, but both their difference and their ratio (rounded to an integer) are—I hope I am safe in concluding—already thousands of digits larger than than any number you have ever counted to.

While there is no precise point at which numbers become unintuitive—it is a gradual process of incomprehension—it's perhaps not that far off the mark to categorize numbers into: small numbers which we can understand distinctly (we, like rabbits, can at a glance tell the difference between 2 people and 3 people), moderately-sized numbers which we have approximate notions (maybe you can quickly tell the difference between a room of 100 people and a room of 1000 people, but perhaps not 800 and 850), big numbers that we only understand in the context of some points of reference (will a billion grains of sand fit into a bucket?), and really huge numbers that we can't intuitively distinguish from infinity (can I imagine so many grains of sand that they could not fit in the known universe? not me). In fact, I mentioned the notion (a minority one) that "actual numbers" may not go on forever. (Note it's commonly believed that there are only finitely many particles in the universe, so we can't "physically" work with numbers beyond a certain point.) In practice, there is not much difference in replacing $\mathbb{N}$ with a model like $\mathbb{L}$ where there are a finite yet sufficiently enormous quantity of numbers that you can still add or multiply any numbers you want to.

I'm sure I haven't satisfactorily argued a reason not to work with $\mathbb{N}$—in fact I don't think there is a good one, so I work with $\mathbb{N}$ all the time. In my understanding there are

two types of concerns. First, that incredibly large numbers have no real physical meaning. I'm somewhat sympathetic to this perspective—maybe really big numbers don't exist in a physical sense, but that's not a problem for working with $\mathbb{N}$ theoretically (i.e., as a theoretically simple model for counting in the physical world). Second, there is a concern (shared, I believe, by a relatively small minority) that because working with infinite sets leads to unintuitive consequences[20], there might be an internal logical paradox if one allows infinitely many numbers. This is something you could possibly be persuaded about if all you know about arithmetic is what your calculator tells you—there turn out to be errors when you work with really large or really small numbers. For instance, the basic calculator on my computer says $0.01^{100} = 0$. A more interesting (and famous) example is if you try to compute $e^{\pi\sqrt{163}}$ on a calculator it looks like an integer: 262537412640768744.0000000000, but you can prove it's not. (That calculation was with 28 digits of precision, but with more, you get 262537412640768743.99999999999925007259....) Since we cannot physically verify the consistency of arithmetic for really large numbers (is $2^{2^{100}} - (2^{2^{100}} - 1)$ really equal to 1 *any* valid way you compute it?), serious skeptics may wonder if there really is some fundamental inconsistency in arithmetic of very large (or very small) numbers akin to the errors a computer may make when doing calculations outside it's usual scope. Still, no one's found one yet, and most people think $\mathbb{N}$ is okay.

Anyway, the point of this course is not to wax philosophical on what is a number, or to defend the use of $\mathbb{N}$ against esoteric skepticism (again, very few people argue against it), but I think it's worthwhile to reflect a little on how and to what extent we actually understand numbers and their arithmetic. For instance, having a good understanding of precision of computations is important if you do any numerical modeling to know how trustworthy your results are. In addition, getting a better sense of the arithmetic of large numbers is important to understand cryptography, which is something we'll touch on when we discuss the RSA cryptosystem in Chapter 3 (e.g., at what point are numbers "too big" to factor)?

Exercise 1.4.4. Go through the five ring axioms, and say which fail for $\mathbb{L}$, and why. Are there any axioms that hold for $\mathbb{N}$ that do not hold for $\mathbb{L}$, or vice versa?

Exercise 1.4.5. Even though $\mathbb{L}$ is closed under addition and multiplication, we've seen that extending $\mathbb{L}$ by including 0 and negative lapine numbers still does not make a ring, unlike when we extend from $\mathbb{N}$ to $\mathbb{Z}$. What property of the addition table of $\mathbb{L}$ prevents this extension of $\mathbb{L}$ from becoming a ring? Why?

1.5 Quadratic rings

After the standard number systems and $\mathbb{Z}/n\mathbb{Z}$, quadratic rings are arguably the next most important and common kinds of number systems in number theory. We already mentioned the Gaussian integers—the set of numbers of the form $a + bi$ where $a, b \in \mathbb{Z}$, in

[20] A striking one for $\mathbb{R}^3$ is the *Banach–Tarski paradox*—look it up, it's amazing. This doesn't mean there's an inherent logical inconsistency in working with $\mathbb{R}^3$, but rather that there are some intuitively incomprehensible consequences of seemingly reasonable assumptions.

the introduction—which will be denoted $\mathbb{Z}[i]$. The idea is the following. For many things, like factoring polynomials or diagonalizing matrices, just working with $\mathbb{R}$ is not sufficient to completely understand things, so one works with $\mathbb{C}$, which is formed by adjoining a formal square root of -1, $i = \sqrt{-1}$, to $\mathbb{R}$. (The number system $\mathbb{C}$ has two square roots of -1—technically we have have to make a choice of one of them to call i, the other will be $-i$. However this choice is not actually important, because if we had made the other choice, the theory is all still the same.)

Similar to this, for many problems in number theory, even though we are often just interested in integer or rational solutions to equations, we can do more things if we consider more general number systems, like adjoining i to $\mathbb{Z}$ to get the Gaussian integers $\mathbb{Z}[i]$. This allows us to factor the expression

$$x^2 + y^2 = (x + iy)(x - iy).$$

Namely, if $n, x, y \in \mathbb{Z}$, then the right hand side of this equation is the product of two Gaussian integers. Thus $\mathbb{Z}[i]$ will be useful dealing with problems where one is led to consider expressions of the form $x^2 + y^2$, such as: find all Pythagorean triples, or what numbers are sums of 2 squares? These questions will be treated in Chapter 4.

More generally, if we have an expression like $x^2 + dy^2$, we can factor this as

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

Here d can be positive or negative.[21] When $d < 0$, looking at expressions of the above form comes up in the classical question: how can you find good rational approximations for square roots? (E.g., the above expression with $d = -2$ is related to rational approximations for $\sqrt{2}$.) This will be the subject of Chapter 5.

In this section, we will introduce quadratic rings[22] and fields, which will be number systems consisting of numbers of the form $a + b\sqrt{d}$, where a, b are either integers or rational numbers. First we give a couple of relevant lemmas. If R is a ring, or $\mathbb{N}$ or $\mathbb{Z}_{\geq 0}$, we say x is a **square** in R if $x = a^2$ for some $a \in R$—otherwise, x is a **non-square**. In particular, $1 = 1^2$ and $0 = 0^2$ are squares in any number system which contains them. Hence a non-square is never 1 nor 0.

Lemma 1.5.1. *Let $d \in \mathbb{Q}$ be a non-square. Then $\sqrt{d} \notin \mathbb{Q}$.*

Proof. (Contrapositive) Suppose $\sqrt{d} \in \mathbb{Q}$. Then we can write $\sqrt{d} = \frac{a}{b}$. Squaring gives $d = (\frac{a}{b})^2$, hence d is a square in $\mathbb{Q}$. □

Probably you've seen a proof that $\sqrt{2}$ is irrational in Discrete Math, and you might remember that being more complicated. Namely, you supposed $\sqrt{2} = \frac{a}{b}$ was rational, multiplied by b and squared both sides, and then gave an argument to get a contradiction.

[21] For any nonzero $d \in \mathbb{Z}$, there are exactly two numbers $z_1, z_2 \in \mathbb{C}$ such that $z_i^2 = d$. So one of these should be $\sqrt{d}$, but we need to make a choice of which one, in order that $\sqrt{d}$ be well defined. We do this using the fact that necessarily $z_2 = -z_1$. If $d > 0$, then each z_i is real, and $\sqrt{d}$ is defined to be the one which is positive. If $d < 0$, then each $z_j = y_j i$ for some $y_j \in \mathbb{R}$, and we take the convention that $\sqrt{d}$ is the z_j such that $y_j > 0$. E.g., when $d = 2$, we take $\sqrt{-2} = \sqrt{2}$. In this way $\sqrt{d}$ is a uniquely defined element of $\mathbb{C}$ for all d. Note this convention agrees with writing $i = \sqrt{-1}$.

[22] Though we will not introduce general quadratic rings yet, just the "naive" ones.

Why was the above so easy? Well, the above lemma is tautological (non-square literally means that a square root does not exist in the ring), and it doesn't actually tell you that $\sqrt{2}$ is irrational because it doesn't tell you 2 is not a square in $\mathbb{Q}$, whereas the standard proof of irrationality of $\sqrt{2}$ does. The proof of irrationality of $\sqrt{2}$ is contained in the following more general result.

Lemma 1.5.2. *Any non-square in $\mathbb{N}$ is a non-square in $\mathbb{Q}$. Hence if $d \in \mathbb{N}$ is a non-square, then $\sqrt{d}$ is irrational. More generally, if $d \in \mathbb{Z}$ is a non-square, then $\sqrt{d}$ is not rational.*

Note the difference between the terms "irrational" and "not rational." Irrational means real but not rational, where as not rational applies to complex numbers as well.

Proof. Let $d \in \mathbb{N}$ be a non-square (meaning it is not a square of an integer). We want to show d is not a square of a rational number. Suppose, for the sake of contradiction, that it is, i.e., $d = (\frac{a}{b})^2$ for some $\frac{a}{b} \in \mathbb{Q}$. We may assume $\frac{a}{b}$ is in reduced form, i.e., a and b have no common prime factors. Then clearing the denominator gives

$$db^2 = a^2.$$

Now let p be a prime factor of d. By the above equation, $p \mid a$. By assumption on $\frac{a}{b}$, $p \nmid b$. Hence the above equation means that p must occur exactly twice as many times in the prime factorization of d as it does for a (e.g., if $p = 2$ and $4 \mid a$ but $8 \nmid a$, then $16 \mid d$ but $32 \nmid a$). Hence the prime(-power) factorization of d looks like

$$d = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r},$$

where each e_i is even. Hence $d = (p_1^{f_1} p_2^{f_2} \cdots p_r^{f_r})^2$, where $f_i = \frac{e_i}{2} \in \mathbb{N}$ for each $1 \leq i \leq r$. Thus d is a square in $\mathbb{N}$, a contradiction. This proves the first statement of the lemma.

The second statement, about $\sqrt{d}$ being irrational, follows from the previous statement together with the previous lemma.

The final statement, when $d \in \mathbb{Z}$ is a non-square, reduces to one of two cases $d \in \mathbb{N}$ or $d \notin N$. Assume d is not the square of an integer. If $d \in \mathbb{N}$, then d cannot be the square of a natural number (if $d = a^2$ for some $a \in \mathbb{Z}$, then $a \neq 0$ so either $\pm a \in \mathbb{N}$, and we can also write $d = (\pm a)^2$), so the previous case applies. So suppose $d \notin \mathbb{N}$. Then d being non-square implies $d \neq 0$, hence $d < 0$. Since all squares in $\mathbb{Q}$ are ≥ 0, d must then be a non-square in $\mathbb{Q}$ and we can apply the previous lemma. □

Exercise 1.5.1. I said the above result contains the proof of the irrationality of $\sqrt{2}$, but to prove this formally, you still need to prove one obvious fact: show 2 is not a square in $\mathbb{N}$. (*One approach:* Try contradiction and think about the absolute value.)

Exercise 1.5.2. Let $n \in \mathbb{Z}$. Prove that n is a square in $\mathbb{Z}$ if and only if it is a square in $\mathbb{Q}$. Is the same true if we replace $\mathbb{Q}$ by $\mathbb{R}$?

Definition 1.5.3. *Let $d \in \mathbb{Z}$ be a non-square. We define the* **quadratic ring**

$$\mathbb{Z}[\sqrt{d}] = \left\{ a + b\sqrt{d} : a, b \in \mathbb{Z} \right\}$$

and the **quadratic field**

$$\mathbb{Q}(\sqrt{d}) = \left\{ a + b\sqrt{d} : a, b \in \mathbb{Q} \right\}.$$

If $d > 0$, we call these rings **real quadratic, and if $d < 0$, we call these imaginary quadratic.**

The condition that d be a non-square is of course so that $\sqrt{d}$ does not already lie in $\mathbb{Z}$ or $\mathbb{Q}$.

The terminology real quadratic versus imaginary quadratic should be self-explanatory. If $d > 0$, then $\sqrt{d}$ is real, so the real quadratic rings and field are contained in $\mathbb{R}$, whereas the imaginary ones are not. Note if $d < 0$, we can write $d = -|d|$, so $\sqrt{d} = \sqrt{|d|}i$ by our standard convention. In this case, we often write elements of $\mathbb{Z}[\sqrt{d}] = \mathbb{Z}[\sqrt{|d|}i]$ and $\mathbb{Q}(\sqrt{d}) = \mathbb{Q}(\sqrt{|d|}i)$ in the form $a + b\sqrt{|d|}i$.

We remark that use of square brackets is standard for rings, and the use of round brackets (parentheses) is standard for fields. However, we often read these two notations the same way, namely "$\mathbb{Z}$ adjoin $\sqrt{d}$" and "$\mathbb{Q}$ adjoin $\sqrt{d}$". (Sometimes people say "bracket" instead of "adjoin.") The idea is that $\mathbb{Z}[\sqrt{d}]$ is the ring you get by adding (adjoining) a square root of d to $\mathbb{Z}$, and similarly $\mathbb{Q}(\sqrt{d})$ is the smallest field by adding a square root of d to $\mathbb{Q}$. Without explaining the technical differences of the notation between square and round brackets in a more general algebraic setting, let me just note that (in this case) $\mathbb{Q}[\sqrt{d}] = \mathbb{Q}(\sqrt{d})$ but $\mathbb{Z}(\sqrt{d}) \neq \mathbb{Z}[\sqrt{d}]$, so you can write $\mathbb{Q}[\sqrt{d}]$ if you really want to, but please don't write $\mathbb{Z}(\sqrt{d})$.

Proposition 1.5.4. *For $d \in \mathbb{Z}$ a non-square, $\mathbb{Z}[\sqrt{d}]$ is a ring and $\mathbb{Q}(\sqrt{d})$ is a field.*

Proof. By Lemma 1.2.6, it suffices to show $\mathbb{Z}[\sqrt{d}]$ and $\mathbb{Q}(\sqrt{d})$ are closed under $+, -$ and $\times$, and that $\mathbb{Q}(\sqrt{d})$ is also closed under division by nonzero elements.

Let us first consider $\mathbb{Z}[\sqrt{d}]$. Consider two elements $a + b\sqrt{d}$ and $a' + b'\sqrt{d}$ in $\mathbb{Z}[\sqrt{d}]$. Then their sum is $(a + a') + (b + b')\sqrt{d}$, their difference is $(a - a') + (b - b')\sqrt{d}$, and their product is $(aa' + dbb') + (ab' + a'b)\sqrt{d}$. These all lie in $\mathbb{Z}[\sqrt{d}]$, hence $\mathbb{Z}[\sqrt{d}]$ is closed under $+, -$ and $\times$.

Now consider $\mathbb{Q}(\sqrt{d})$. By the same argument as for $\mathbb{Z}[\sqrt{d}]$, $\mathbb{Q}(\sqrt{d})$ is closed under $+, -$ and $\times$. We need to show it is closed under division by nonzero elements, i.e., $\alpha/\beta \in \mathbb{Q}(\sqrt{d})$ for all $\alpha, \beta \in \mathbb{Q}(\sqrt{d})$ with $\beta \neq 0$. Since we already know $\mathbb{Q}(\sqrt{d})$ is closed under multiplication, rewriting $\alpha/\beta = \alpha \cdot 1/\beta$ (valid as complex numbers), it suffices to show any nonzero $\beta \in \mathbb{Q}(\sqrt{d})$ has a multiplicative inverse, i.e., $1/\beta \in \mathbb{Q}(\sqrt{d})$.

Say $\beta = a + b\sqrt{d} \in \mathbb{Q}(\sqrt{d})$ is nonzero. We want to say there exists $a' + b'\sqrt{d} \in \mathbb{Q}(\sqrt{d})$ such that $1/\beta = a' + b'\sqrt{d}$, i.e.,

$$(a + b\sqrt{d})(a' + b'\sqrt{d}) = (aa' + dbb') + (ab' + a'b)\sqrt{d} = 1.$$

If $b = 0$, we can simply take $a' + b'\sqrt{d} = 1/a$, so assume $b \neq 0$. Then setting $a' = -\frac{ab'}{b}$ and $b' = (b(d - (a/b)^2))^{-1}$ gives the desired equality. Note that a' and b' are both well defined by the assumptions that $b \neq 0$ and d is a non-square in $\mathbb{Z}$, whence $d - (a/b)^2 \neq 0$ by the above lemma. □

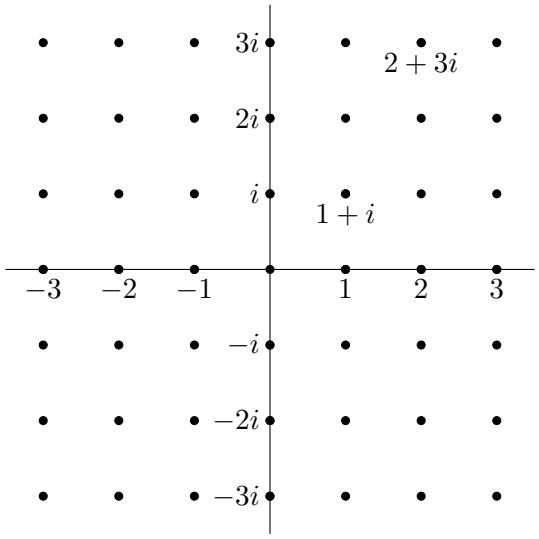


Figure 1.5.1: $\mathbb{Z}[i]$ inside $\mathbb{C}$

It is important in the above definition of $\mathbb{Z}[\sqrt{d}]$ that we took d to be an integer, as the following shows.

Exercise 1.5.3. Let $d = \frac{1}{2}$. Show $\left\{a + b\sqrt{d} : a, b \in \mathbb{Z}\right\}$ is not a ring, though $\left\{a + b\sqrt{d} : a, b \in \mathbb{Q}\right\}$ is a field.

We can think of the ring $\mathbb{Z}[\sqrt{d}]$ inside $\mathbb{Q}(\sqrt{d})$ as being analogous to $\mathbb{Z}$ inside $\mathbb{Q}$—namely $\mathbb{Q}(\sqrt{d})$ is the field obtained by taking ratios of two elements of $\mathbb{Z}[\sqrt{d}]$, and $\mathbb{Z}[\sqrt{d}]$ as being like integers. Consequently, we will call elements of $\mathbb{Z}[\sqrt{d}]$ **quadratic integers**, though there are other numbers that are considered quadratic integers as well, e.g., $\frac{1+\sqrt{-3}}{2}$. We won't get into why $\frac{1+\sqrt{-3}}{2}$ should be considered as an "integer" now, but we'll see this particular number in the next section.

First let's take a look at some imaginary quadratic examples.

Example 1.5.1. Let $d = -1$. Then $\mathbb{Z}[\sqrt{-1}] = \mathbb{Z}[i]$ is the ring of **Gaussian integers**, and $\mathbb{Q}(\sqrt{-1}) = \mathbb{Q}(i)$ is the field of **Gaussian numbers**, which we saw briefly in the introduction. Just like we drew $\mathbb{Z}$ on the real number line in Fig. 1.1.1, we can draw $\mathbb{Z}[i]$ on the complex plane as in Fig. 1.5.1.

Example 1.5.2. For $d = -3$, $\sqrt{d} = \sqrt{-3} = \sqrt{3}i$, and we can draw the ring $\mathbb{Z}[\sqrt{-3}] = \left\{a + \sqrt{3}bi : a, b \in \mathbb{Z}\right\}$ in $\mathbb{C}$ as in Fig. 1.5.2. Note this looks like the picture for $\mathbb{Z}[i]$, just scaled out vertically by a factor of $\sqrt{3}$.

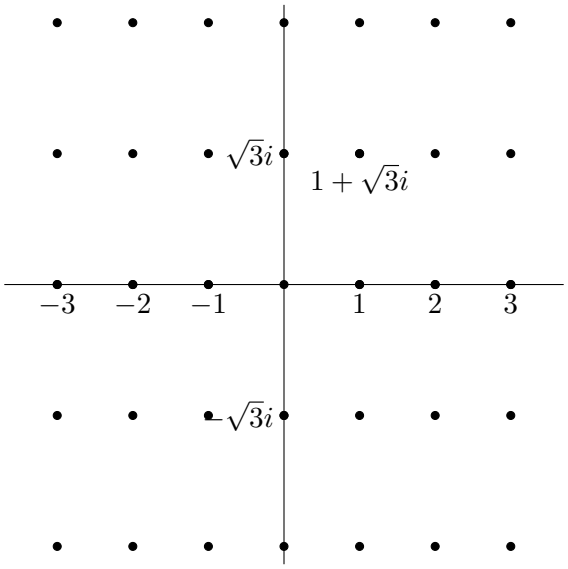


Figure 1.5.2: $\mathbb{Z}[\sqrt{-3}]$ inside $\mathbb{C}$

Example 1.5.3. Let $d = -4$. Then $\mathbb{Z}[\sqrt{d}] = \mathbb{Z}[\sqrt{-4}i] = \mathbb{Z}[2i] = \{a + 2bi : a, b \in \mathbb{Z}\}$. Note that this is a subring of the Gaussian integers $\mathbb{Z}[i]$, and we can visualize it as the subset of $\mathbb{Z}[i]$ by removing every other row of dots in Fig. 1.5.1. Clearly, this is a proper subring, i.e., $\mathbb{Z}[2i] \neq \mathbb{Z}[i]$, because, for instance, the Gaussian integer $i \notin \mathbb{Z}[2i]$.

On the other hand, we claim that $\mathbb{Q}(\sqrt{d}) = \mathbb{Q}(2i) = \mathbb{Q}(i)$. First, given any $a + 2bi \in \mathbb{Q}(2i)$ (so $a, b \in \mathbb{Q}$), we can write this as $a + b'i \in \mathbb{Q}(i)$ with $b' = 2b \in \mathbb{Q}$. Hence $\mathbb{Q}(2i) \subset \mathbb{Q}(i)$. Conversely, if $a + bi \in \mathbb{Q}(i)$, then we can rewrite this as $a + 2b'i \in \mathbb{Q}(2i)$ where $b' = \frac{b}{2} \in \mathbb{Q}$. Thus $\mathbb{Q}(i) \subset \mathbb{Q}(2i)$, and these sets are equal.

Generalizing the previous example, are a few exercises about how quadratic rings and fields are related for different choices of d.

Exercise 1.5.4. Let $d, d' \in \mathbb{Z}$ be non-squares. Show that $\mathbb{Z}[\sqrt{d'}]$ is a subring of $\mathbb{Z}[\sqrt{d}]$ if and only if $d' = n^2 d$ for some $n \in \mathbb{N}$. Under this condition, when will $\mathbb{Z}[\sqrt{d'}]$ be a *proper* subring of $\mathbb{Z}[\sqrt{d}]$, i.e., a subring of $\mathbb{Z}[\sqrt{d}]$ which is not equal to $\mathbb{Z}[\sqrt{d}]$?

Exercise 1.5.5. Let $d, d' \in \mathbb{Z}$ be non-squares. Show $\mathbb{Z}[\sqrt{d}] \subset \mathbb{Z}[\sqrt{d'}]$ implies $\mathbb{Q}(\sqrt{d}) = \mathbb{Q}(\sqrt{d'})$.

Exercise 1.5.6. Find an example of non-squares $d, d' \in \mathbb{Z}$ such that $\mathbb{Q}(\sqrt{d}) = \mathbb{Q}(\sqrt{d'})$ but neither $d \mid d'$ nor $d' \mid d$ is true.

Now let's look at a real quadratic example.

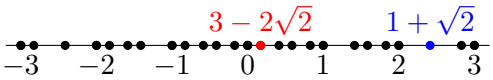


Figure 1.5.3: A sample of $\mathbb{Z}[\sqrt{2}]$ inside $\mathbb{R}$

Example 1.5.4. Let $d = \sqrt{2}$. Then $\mathbb{Z}[\sqrt{2}] = \{a + b\sqrt{2} : a, b \in \mathbb{Z}\}$ and $\mathbb{Q}(\sqrt{2})$ are contained in $\mathbb{R}$. We can draw these as point on the real line, however the picture will look very different than than imaginary quadratic integers, or of $\mathbb{Z}$. For those situations the picture of integers is what is called a *lattice*—in particular points are well spaced out, so there are only finitely many points within a finite region of the plane (in the case of imaginary quadratic integers) or the line (in the case of $\mathbb{Z}$). However, the more elements of $\mathbb{Z}[\sqrt{2}]$ we draw (say, draw $a + b\sqrt{2}$, with $|a|, |b| < N$ for some N, and then do this for larger and larger N), we'll see that points are getting closer and closer together.

See Fig. 1.5.3 for a picture of all $a + b\sqrt{2}$, with $|a|, |b| \leq 3$, which lie between -3 and 3.

We can formally state the difference between the pictures for imaginary and real quadratic integers in the following.

Proposition 1.5.5. *Let $d \in \mathbb{Z}$ be a non-square.*

(1) (Imaginary quadratic case) Suppose $d < 0$. Then $\mathbb{Z}[\sqrt{d}]$ is a discrete *subset of $\mathbb{C}$, i.e., there are only finitely many elements of $\mathbb{Z}[\sqrt{d}]$ within any bounded region (e.g., a rectangle or a circle) in the complex plane.*

(2) (Real quadratic case) Suppose $d > 0$. Then $\mathbb{Z}[\sqrt{d}]$ is a dense *subset or $\mathbb{R}$, i.e., there exists an element of $\mathbb{Z}[\sqrt{d}]$ (in fact infinitely many) inside any non-empty open interval (x_1, x_2) of $\mathbb{R}$.*

We won't prove the real quadratic case (which is not super important for this class anyway), but the essential aspects of the proof are contained in the following special case:

Exercise 1.5.7. Show that for any $\varepsilon > 0$, there exists an element $a + b\sqrt{2}$ of $\mathbb{Z}[\sqrt{2}]$ in the interval $(0, \varepsilon)$. Use this to conclude that there are infinitely many elements of $\mathbb{Z}[\sqrt{2}]$ close to 0—specifically, for any $\varepsilon > 0$, there are infinitely many elements of $\mathbb{Z}[\sqrt{2}]$ in $(0, \varepsilon)$. (*Suggestion:* Think about the decimal expansion of $\sqrt{2}$.)

For the imaginary quadratic case, it will be convenient to use the following fundamental concept from algebraic number theory.

Definition 1.5.6. *Let $d \in \mathbb{Z}$ be a non-square. For $\alpha = a + b\sqrt{d} \in \mathbb{Q}(\sqrt{d})$, we define the* **conjugate** *of α to be*

$$\overline{\alpha} = a - b\sqrt{d}.$$

The **norm** *of α is defined by*

$$N(a + b\sqrt{d}) = N(\alpha) = \alpha\overline{\alpha} = (a + b\sqrt{d})(a - b\sqrt{d}) = a^2 - db^2.$$

Note for $a, b \in \mathbb{Q}$, $N(a+b\sqrt{d}) \in \mathbb{Q}$, so we can think of the norm as a map $N : \mathbb{Q}(\sqrt{d}) \to \mathbb{Q}$. Similarly, if $a, b \in \mathbb{Z}$, then $N(a + b\sqrt{d}) = a^2 - db^2 \in \mathbb{Z}$, so the norm of a quadratic integer is an ordinary integer, i.e., $N : \mathbb{Z}[\sqrt{d}] \to \mathbb{Z}$.

You can think of the norm an algebraic way of measuring the "size" of a quadratic number. In the imaginary quadratic case, it already corresponds to a geometric notion you know: Say $d < 0$. If we think of $z = a+b\sqrt{d} = a+b\sqrt{|d|}i$ as a vector in $\mathbb{C} \simeq \mathbb{R}^2$, it is a vector with length $\sqrt{a^2 + |d|b^2}$, i.e., the $N(a+b\sqrt{d})$ is the *square* of the length of z. Alternatively, we can define the ordinary complex absolute value $|z|$ for any $z \in \mathbb{C}$ by $|z| = \sqrt{z\overline{z}}$, where $\overline{z}$ denotes *complex* conjugation. In the imaginary quadratic case, the conjugation we defined above agrees with complex conjugation, and for $z = a + b\sqrt{d}$, $N(z) = z\overline{z} = |z|^2$. For arithmetic purposes, it is better to work with the norm than the usual absolute value (for instance, so the norm of a quadratic integer is an integer).

In the real quadratic case, we don't have the same interpretation, but the above algebraic definition of norm makes equal sense in the imaginary and real quadratic settings. So you can think of the norm in the real quadratic case as an alternative, more arithmetic, measure of size than the usual absolute value. Note one big difference between the imaginary and real quadratic cases: in the imaginary quadratic case the norm map is always non-negative, but in the real quadratic case the norm takes on both positive and negative values. For instance, the highlighted points in Fig. 1.5.3 have norms $N(3 - 2\sqrt{2}) = 9 - 2 \cdot 4 = 1$ and $N(1 + \sqrt{2}) = 1^2 - 2 \cdot 1^2 = -1$ (note also $N(1) = N(-1) = 1$). While there is no apparent relation between the norm of real quadratic number and where it lies on the real line, the norm is still an algebraically useful quantity to look at.

Proof of Proposition in imaginary quadratic case. Let $d < 0$. We want to show that any bounded region in $\mathbb{C}$ contains only finitely many elements of $\mathbb{Z}[\sqrt{d}]$. Any bounded region in $\mathbb{C}$ must lie within an ellipse of the form

$$E_n = \{x + iy : x, y \in \mathbb{R},\, x^2 + |d|y^2 \le n\}$$

for large enough n. Now the elements of $\mathbb{Z}[\sqrt{d}]$ which lie in E_n are precisely the elements $a+b\sqrt{d}$ of $\mathbb{Z}[\sqrt{d}]$ with norm up to n. But if $N(a+b\sqrt{d}) = a^2+|d|b^2 \le n$ then necessarily $|a| \le \sqrt{n}$ and $|b| \le \sqrt{n}$ (in fact $|b| \le \sqrt{\frac{n}{|d|}}$), i.e., we must have $a, b \in \{-n, -(n-1), \dots, n-1, n\}$. Hence there are at most $(2n+1)^2$ elements of $\mathbb{Z}[\sqrt{d}]$ in E_n. □

There are two main properties of the norm that make it very useful (in both the real and imaginary settings): (1) it takes quadratic integers to integers, and (2) it has the following multiplicativity property:

Exercise 1.5.8. Let $d \in \mathbb{Z}$ be a non-square. For $\alpha, \beta \in \mathbb{Q}(\sqrt{d})$, show that $N(\alpha\beta) = N(\alpha)N(\beta)$.

We will exploit these properties of the norm in later chapters. As a teaser, if we want to know what numbers n are the sum of two (integer) squares, that means determining n for which $a^2 + b^2 = n$ has a solution in $\mathbb{Z}$, i.e., the n for which there is a Gaussian integer

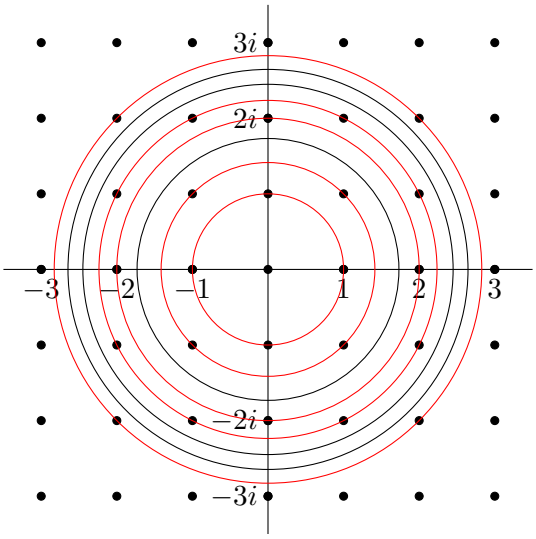


Figure 1.5.4: $\mathbb{Z}[i]$ with circles of radius $\sqrt{n}$, $1 \le n \le 8$

$z = a + bi$ of norm n. Since $N(z) = \sqrt{|z|}$, this means an integer n is the sum of two squares if and only if the circle of radius $\sqrt{n}$ centered at 0 intersects a Gaussian integer.

In Fig. 1.5.4, I've drawn the circles of radius $\sqrt{n}$, $1 \le n \le 8$ on top of our picture of $\mathbb{Z}[i]$, and highlighted in red the ones that hit Gaussian integers. In particular, we see 1, 2, 4, 5 and 8 are sums of two squares while 3, 6 and 7 are not.

1.6 Cyclotomic rings

Besides the standard number systems, the main ones we will use in this course are $\mathbb{Z}/n\mathbb{Z}$, $\mathbb{Z}[\sqrt{d}]$ and $\mathbb{Q}(\sqrt{d})$. However, there are a couple of other ones that will come up. Here we will briefly introduce cyclotomic rings. Whereas working with quadratic rings allows us to factor quantities of the form $x^2 + dy^2$, cyclotomic rings will allow us to factor quantities of the form $x^n + y^n$, and thus are relevant for Fermat's last theorem.

To introduce cyclotomic rings, we first need to introduce roots of unity, which are a beautiful piece of mathematics all math majors should be familiar with.

Definition 1.6.1. *Let $n \in \mathbb{N}$. The n-th roots of unity are the elements $z \in \mathbb{C}$ such that $z^n = 1$. We denote the set of n-th roots of unity by μ_n.*

The simplest cases which you should already be familiar with are: $\mu_1 = \{1\}$, $\mu_2 = \{1, -1\}$, $\mu_4 = \{1, -1, i, -i\}$.

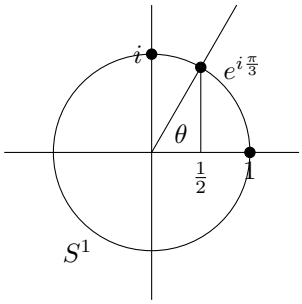
Exercise 1.6.1. Show that if $m \mid n$, $\mu_m \subset \mu_n$.

The way to determine μ_n in general comes from using polar form for complex numbers. Namely, we can write any $z \in \mathbb{C}$ in the form $z = re^{i\theta}$ where $r, \theta \in \mathbb{R}$ with $r \ge 0$. Here

$$e^{i\theta} = \cos\theta + i \sin\theta$$

(you can take this as the definition of $e^{i\theta}$ if you have complex exponential and trig functions before). Let S^1 denote the circle of radius 1 centered at 0. Since $\cos^2\theta + \sin^2\theta = 1$, $e^{i\theta}$ is the point on the circle S^1 which lies on a ray through the origin at angle θ from the positive real axis.

Example 1.6.1. If we take $\theta = \frac{\pi}{3}$, then $\cos\theta = \frac{1}{2}$ and $\sin\theta = \frac{\sqrt{3}}{2}$ so $e^{i\frac{\pi}{3}} = \frac{1+\sqrt{3}i}{2} \in \mathbb{Q}(\sqrt{-3})$.



Now the polar form is not unique, but if $z = re^{i\theta}$ is not zero, it is unique if we require $\theta \in [0, 2\pi)$. Here r tells us the distance z is from the origin, i.e., $r = |z|$ and θ tells us on what ray through the origin z lies on.

Multiplication of real numbers has a geometric interpretation: if $r > 0$, multiplication by r effects scaling the real line by $\mathbb{R}$, and if $r < 0$, multiplication by r is reflection about 0 composed with scaling by $|r|$. So too does multiplication of complex numbers, which is easiest seen from the polar form $re^{i\theta}$. Multiplication by $z = re^{i\theta}$ scales radially outward by r and rotates about 0 by θ. To see this, take some $w \in \mathbb{C}$ which we write in polar form as $w = se^{i\phi}$. Then

$$zw = re^{i\theta}se^{i\phi} = rse^{i(\theta+\phi)}.$$

Now let's suppose $z = re^{i\theta} \in \mu_n$, i.e., $z^n = 1$. Assume $0 \le \theta < 2\pi$ so this representation is unique. Then

$$z^n = r^n e^{in\theta} = 1 \implies r = 1,\ n\theta \in 2\pi\mathbb{Z}.$$

That is z must be one of the n following numbers

$$1 = e^{i0}, e^{i\frac{2\pi}{n}}, e^{i\frac{4\pi}{n}}, e^{i\frac{6\pi}{n}}, \dots, e^{i\frac{2\pi(n-1)}{n}}.$$

Furthermore, these all lie in μ_n so they are precisely the n-roots of unity. (Here is another reason there should be n elements of μ_n for all n: each $z \in \mu_n$ corresponds to a root of the polynomial $x^n - 1$, which must have n roots by the fundamental theorem of algebra.)

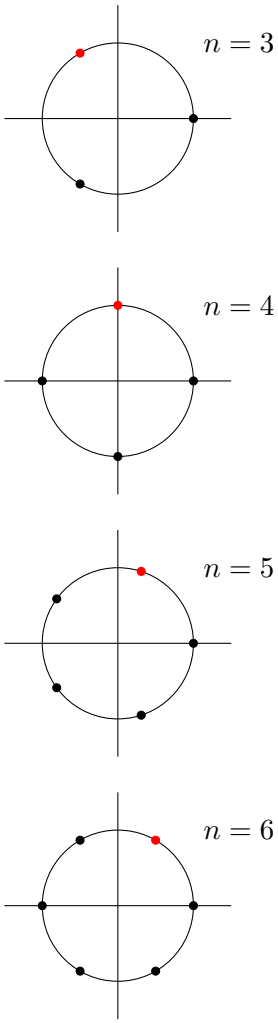
We will denote the first nontrivial solution on this list by:

$$\zeta_n = e^{\frac{2\pi i}{n}}.$$

Then we can write the n-th roots of unity as

$$\mu_n = \left\{ e^{\frac{2\pi k i}{n}} : 0 \le k < n \right\} = \left\{1, \zeta_n, \zeta_n^2, \dots, \zeta_n^{n-1}\right\}.$$

One beautiful thing about the n-th roots of unity is they are the vertices of a regular n-gon inscribed in S^1. Here are a few pictures. (You can play connect-the-dots yourself to see a regular n-gon.) In each case ζ_n is the root of unity in red. (The sequence $1, \zeta_n, \zeta_n^2, \dots, \zeta_n^{n-1}$ goes in counterclockwise order, as should be clear from the geometry of multiplication: multiplication by ζ_n simply acts as rotation by $\frac{2\pi}{n}$.)



The cyclotomic rings are the rings that are generated by these roots of unity.

Definition 1.6.2. *The* n**-th cyclotomic ring (of integers)** *is*

$$\mathbb{Z}[\zeta_n] = \left\{a_0 + a_1\zeta_n + a_2\zeta_n^2 + \cdots + a_{n-1}\zeta_n^{n-1} : a_i \in \mathbb{Z}\right\},$$

and the n**-th cyclotomic field** *is*

$$\mathbb{Q}(\zeta_n) = \left\{a_0 + a_1\zeta_n + a_2\zeta_n^2 + \cdots + a_{n-1}\zeta_n^{n-1} : a_i \in \mathbb{Q}\right\}.$$

Exercise 1.6.2. Prove that $\mathbb{Z}[\zeta_n]$ is a ring.

Exercise 1.6.3. Prove that $\mathbb{Q}(\zeta_n)$ is a field.

When $n = 1$, $\zeta_1 = 1$, so $\mathbb{Z}[\zeta_1] = \mathbb{Z}$ and $\mathbb{Q}(\zeta_1) = \mathbb{Q}$. When $n = 2$, $\zeta_2 = -1$, so $\mathbb{Z}[\zeta_2] = \{a_0 + a_1(-1) : a_1 \in \mathbb{Z}\} = \mathbb{Z}$ and $\mathbb{Q}(\zeta_2) = \mathbb{Q}$.

We note that for $n > 1$, unlike the case of quadratic ring, the representation of a cyclotomic number as $a_0 + a_1\zeta_n + a_2\zeta_n^2 + \cdots + a_{n-1}\zeta_n^{n-1}$ is *not* unique, i.e. there are $\mathbb{Z}$-linear relations between $1, \zeta_n, \ldots, \zeta_n^{n-1}$, i.e., $1, \zeta_n, \ldots, \zeta_n^{n-1}$ is not a "basis." For instance when $n = 2$ we have the relation $1+\zeta_2 = 1+(-1) = 0$, and when $n = 4$ one has $\zeta_4+\zeta_4^2 = i+i^3 = 0$.

Example 1.6.2. When $n = 4$, we have $\zeta_4 = i$, and $\mathbb{Z}[\zeta_4] = \{a_0 + a_1 i + a_2(-1) + a_3(-i) : a_i \in \mathbb{Z}\} = \mathbb{Z}[i]$. Similarly $\mathbb{Q}(\zeta_4) = \mathbb{Z}[\zeta_4]$.

Example 1.6.3. When $n = 6$, we see from Example 1.6.1 that $\zeta_6 = \frac{1+\sqrt{3}i}{2}$. Note $\zeta_6^2 = \zeta_3 = \frac{-1+\sqrt{3}i}{2} = \zeta_6 - 1$, $\zeta_6^3 = \zeta_2 = -1$, $\zeta_6^4 = \zeta_6^3\zeta_6 = -\zeta_6$, and $\zeta_6^5 = \zeta_6^3\zeta_6^2 = -\zeta_3$. Consequently, we can write all powers of ζ_6 as integer combinations of either ζ_3 or ζ_6, and we see that we can simply write all cyclotomic integers for $n = 3$ or $n = 6$ as

$$\mathbb{Z}[\zeta_3] = \mathbb{Z}[\zeta_6] = \{a + b\zeta_6 : a, b \in \mathbb{Z}\} = \left\{a + b\frac{1+\sqrt{-3}}{2} : a, b \in \mathbb{Z}\right\}.$$

Note that $\mathbb{Z}[\zeta_3] \subset \mathbb{Q}(\sqrt{-3})$ but it is not contained in the quadratic ring $\mathbb{Z}[\sqrt{-3}]$. We call $\mathbb{Z}[\zeta_3]$ the **Eisenstein integers** (named in honor of FGM Eisenstein, who died of TB at 29).

A brief digression about quadratic rings: It turns out that sometimes the set of numbers of the form $a + b\frac{1+\sqrt{d}}{2}$ $(a, b \in \mathbb{Z})$ behaves better than $\mathbb{Z}[\sqrt{d}]$. This is the case for $d = -3$, where we get the Eisenstein integers. They will not always form a ring, but when they do, we consider elements of this form to be quadratic integers as well. In particular, we consider $\mathbb{Z}[\zeta_3]$ be a quadratic ring of integers. Here is a real quadratic example:

Exercise 1.6.4. Let $\phi = \frac{1+\sqrt{5}}{2}$ be the golden ratio. Show $\mathbb{Z}[\phi] = \{a + b\phi : a, b \in \mathbb{Z}\}$ is a subring of $\mathbb{Q}(\sqrt{5})$.

Just to see what can go wrong with numbers of this form:

Exercise 1.6.5. Show $\left\{a + b\frac{1+\sqrt{3}}{2} : a, b \in \mathbb{Z}\right\}$ is not a ring.

We remark that there is an elementary criterion for when the set of numbers of the form $a + b\frac{1+\sqrt{d}}{2}$ $(a, b \in \mathbb{Z})$ is a ring: it happens exactly when $d \equiv 1 \bmod 4$.

1.7 Beyond $\mathbb{C}$

All of the number rings we looked at in this chapter beyond $\mathbb{Z}/n\mathbb{Z}$ were subrings of $\mathbb{C}$. You might wonder if there are other interesting kinds of numbers not contained in $\mathbb{C}$. Indeed there are.

One type of example is given by the **p-adic integers** $\mathbb{Z}_p$ and the **p-adic numbers** $\mathbb{Q}_p$, where p is a prime number. The basic idea is we can write any positive integer n in base p:

$$n = a_0 + a_1 p + a_2 p^2 + \cdots + a_r p^r, \quad 0 \le a_i < p,$$

Instead of just working with finite p-adic expansions, we work with *infinite* ones:

$$a_0 + a_1 p + a_2 p^2 + \cdots, \quad 0 \le a_i < p.$$

Here the sum diverges, but it is not meant to be evaluated, it is meant to be thought of a limit of a base p-expansion of an integer. You can add and multiply them subject to the usual rules, and you can even subtract them. For instance, if $p = 3$, -2 is given by

$$1 + 2p + 2p^2 + 2p^3 + 2p^4 + \cdots$$

(Just add 2, and do the carry overs.) The set of such formal infinite series is $\mathbb{Z}_p$.

It's less obvious, but you can also divide them (most of the time): for instance again with $p = 3$, $1/2$ is given by

$$2 + p + p^2 + p^3 + p^4 + \cdots$$

(Multiply by 2 and do the carry overs.) More precisely, you can divide $\sum a_i p^i$ by $\sum b_i p^i$ when $b_0 \neq 0$. To take general quotients, you need to work with formal Laurent series, i.e., expressions of the form

$$a_{-r} p^{-r} + a_{1-r} p^{-r} + \cdots + a_0 + a_1 p + a_2 p^2 + a_3 p^3 + \cdots, \quad 0 \le a_i < p.$$

Elements of this form give you a field, $\mathbb{Q}_p$.

We won't work with p-adic numbers in this class, but they're a convenient way to study the all rings $\mathbb{Z}/p\mathbb{Z}$, $\mathbb{Z}/p^2\mathbb{Z}$, $\mathbb{Z}/p^3\mathbb{Z}$, ..., simultaneously, and are incredibly important in more advanced number theory.[23]

Another type of number system is given by the **quaternions**. The idea is just like we can describe rotations in the plane (about 0) by multiplication by complex numbers $e^{i\theta}$, William Rowan Hamilton wondered if there is a 3-dimensional type of number system whereby multiplication would realize 3-d rotations. After about 10 years, he realized this was impossible, but you could instead do it in 4-dimensions!

The Hamilton quaternions are given by

$$\mathbb{H} = \{a + bi + cj + dk : a, b, c, d \in \mathbb{R}\},$$

where i, j, and k are quantities such that

$$i^2 = j^2 = k^2 = -1, \quad k = ij = -ji.$$

[23] Technically it is possible to abstractly identify $\mathbb{Z}_p$ and $\mathbb{Q}_p$ with subrings of $\mathbb{C}$ using the axiom of choice, but it's not concrete and doesn't really shed any light on these rings.

You can define addition, and extend the definition of multiplication to make $\mathbb{H}$ into a *noncommutative ring*. (Addition is still commutative, but multiplication is not.) The quaternions were actually a precursor to linear algebra, and actually have some advantages over traditional linear algebra techniques—they are still used in engineering and computing to work with 3-d rotations, being more efficient for calculations than standard matrix representations. (You only need 4 real numbers to represent a quaternion, whereas you need 9 real numbers to represent a 3×3 matrix.) Here the fact that multiplication is noncommutative corresponds to the fact that if you take two 3-d rotations and compose them, the result in general depends on which order you do them in.

In regards to number theory, one can look at integers in $\mathbb{H}$, which can be defined in various ways, but the simplest is just

$$R = \{a + bi + cj + dk : a, b, c, d \in \mathbb{Z}\} \subset \mathbb{H}.$$

Such rings are very useful in number theory as well—for instance you can use quaternions to determine what numbers are sums of 3 or 4 squares. We discuss this to some extent in Chapter 4.

Shortly after Hamilton's discovery of the quaternions, Graves and Cayley discovered even higher-dimensional generalizations like the **octonions**. These are not even associative! However, there is still a fair amount of structure in the octonions (they are "almost associative"), and they also have interesting applications to number theory, but we will not cover them in this course.

Finally, we mention that there are other number systems extending $\mathbb{R}$ to treat both infinitesimal and infinite quantities, such as the **hyperreals** and **surreals**. The idea is that one can do algebra with both infinitesimal and infinite quantities and sometimes get something that seems correct (e.g., multiplying $\frac{dy}{dx}$ by dx). There are number systems that make such infinitesimal and infinite arithmetic formal procedures. Personally, I find these philosophically appealing, but it seems the foundational theory is too difficult at present for these systems to have found widespread use. In any case, such topics are not part of the present course.

Chapter 2

Factorization

In this chapter, we will prove the fundamental theorem of arithmetic, i.e., the uniqueness of prime factorization for natural numbers. However, we will set up the framework for this more generally. This is for two reasons. First, we will want to use uniqueness of prime factorization for some quadratic rings as well, so we want to be able to prove it for the Gaussian integers, for instance. Second, your familiarity with unique factorization makes it harder to appreciate—unique factorization is a nontrivial property and it does not hold for many rings. I hope that putting unique factorization in the context of more general rings—and seeing how it fails for some quadratic rings—may help you appreciate how special it is.

To do this, we need to figure out what the right notion of unique factorization is in general. Let's recall our previous statement of the **fundamental theorem of arithmetic**:

Theorem 2.0.1 (Unique factorization for $\mathbb{N}$)**.** *Let $n > 1$ be a natural number. Then n factors into a product of prime numbers. Moreover, this factorization is unique up to reordering, i.e., if*

$$n = p_1 p_2 \cdots p_r = q_1 q_2 \cdots q_s,$$

where the p_i's and q_j' are primes, and are ordered so that

$$p_1 \leq p_2 \leq \cdots \leq p_r, \quad q_1 \leq q_2 \leq \cdots \leq q_s,$$

then $r = s$ and $p_i = q_i$ for each $1 \leq i \leq r$.

We'd like to talk about factorization in rings, so let's think about how we can restate this for $\mathbb{Z}$. We just need to say any integer n which is not 0 or ± 1 is ± 1 times a product of primes, and the primes in this product are uniquely determined.

Theorem 2.0.2 (Unique factorization for $\mathbb{Z}$)**.** *Let $n \in \mathbb{Z}$ be nonzero and $n \neq \pm 1$. Then we can write $n = up_1 p_2 \cdots p_r$, where $u = \pm 1$ and $p_1, p_2, \ldots, p_r$ are prime. Moreover, this factorization is unique up to reordering, i.e., if*

$$n = up_1 p_2 \cdots p_r = u' q_1 q_2 \cdots q_s,$$

where u and u' are ± 1, and the p_i's and q_j' are primes ordered so that

$$p_1 \leq p_2 \leq \cdots \leq p_r, \quad q_1 \leq q_2 \leq \cdots \leq q_s,$$

then $r = s$ and $p_i = q_i$ for each $1 \leq i \leq r$.

Note that the sign u is also uniquely determined, though we did not state this. Moreover, if we want to, we can actually omit the condition that $n \neq \pm 1$ by allowing r to be 0, i.e., allowing the "factorization" $n = u$.

2.1 Units, irreducibles and existence of factorizations

To generalize the notion of unique factorization from $\mathbb{Z}$ to more general rings R, we need to introduce some terminology. We will primarily be concerned with the cases of $R = \mathbb{Z}$ and R is a quadratic ring. To treat these uniformly, we will define

$$\mathbb{Z}[\sqrt{d}] = \mathbb{Z} \quad \text{if } d \text{ is a square.}$$

E.g., $\mathbb{Z} = \mathbb{Z}[\sqrt{1}]$. This coincides with the notion that $\mathbb{Z}[\sqrt{d}]$ is the ring obtained by adjoining a square root of d to $\mathbb{Z}$—if d is already a square, there is nothing to add. Thus $\mathbb{Z}[\sqrt{d}]$ for $d \in \mathbb{Z}$ will either mean $\mathbb{Z}$ or a quadratic ring $\mathbb{Z}[\sqrt{d}]$ for some non-square d. (Recall, there are other quadratic rings, like $\mathbb{Z}[\zeta_3]$, but for simplicity we will not worry about those now.)

We will also want to talk about the "size" of integers, as we want to think about the factorization of a number as breaking a number into "smallest possible" components. The norm provides a measure of size for quadratic rings, so the norm will be key for us. So that we can talk about the norm N on $\mathbb{Z}[\sqrt{d}]$ in all cases, we simply define the **norm** on $\mathbb{Z}$ to be $N(n) = n$ when $\mathbb{Z}[\sqrt{d}] = \mathbb{Z}$.[1] (We note that one could alternatively take something like $N(n) = |n|$ or $N(n) = n^2$ for what we are going to do, but it is standard to define the norm on $\mathbb{Z}$ to just be the identity map.)

Then, for any $d \in \mathbb{Z}$, we have the following key properties of the norm map N on $\mathbb{Z}[\sqrt{d}]$:

- $N : \mathbb{Z}[\sqrt{d}] \to \mathbb{Z}$, i.e., the norm of any element of $\mathbb{Z}[\sqrt{d}]$ is an integer;
- $N(xy) = N(x)N(y)$, i.e., the norm map is *multiplicative*; and
- $N(x) = 0$ if and only if $x = 0$.

For d non-square, the first property followed directly from the definition. The second property was Exercise 1.5.8. The third is a simple exercise:

> **Exercise 2.1.1.** Let $d \in \mathbb{Z}$ be a non-square. For $x \in \mathbb{Z}[\sqrt{d}]$, show $N(x) = 0$ if and only if $x = 0$.

Note when d is a square, so $\mathbb{Z}[d] = \mathbb{Z}$ and $N(x) = x$, all 3 properties are obvious.

While I often say the norm measures the size of a number, you are probably use to thinking of size as a positive quantity. Since the norm may be negative, we will sometimes work with the **absolute norm** $|N(x)|$. This is again multiplicative, but now $|N(x)| \in \mathbb{N}$ for any $x \neq 0$. In particular, the absolute norm will allow us to use descent, and thus prove the existence of factorizations.

[1] Note that for d not a square in $\mathbb{Z}$, the norm map from $\mathbb{Z}[\sqrt{d}]$ to $\mathbb{R}$ sends any $n \in \mathbb{Z}$ to n^2. So when we talk about the norm of an integer, it is important to know whether we mean the norm from $\mathbb{Z}$ or the norm from a quadratic ring $\mathbb{Z}[\sqrt{d}]$.

While we will primarily prove things for quadratic rings (and $\mathbb{Z}$), many of the definitions we will state for more general rings. This way, we can talk about these notions for other rings like cyclotomic rings as well.[2] First, we need the analogue of ± 1 in $\mathbb{Z}$.

Definition 2.1.1. *Let R be a ring and $u \in R$. We say u is a* **unit** *in R if u is invertible in R, i.e., if there exists $u^{-1} \in R$ such that $uu^{-1} = 1$.*

Example 2.1.1. Let $R = \mathbb{Z}$ and $n \in \mathbb{Z}$. Let's prove that n is invertible in $\mathbb{Z}$ if and only if $n = \pm 1$. First note that 0 is not invertible, since $0 \cdot m = 0 \neq 1$ for all $m \in \mathbb{Z}$. Then if $n \neq \pm 1$ and $n \neq 0$, $|n| \geq 2$ so $|nm| \geq 2$ for any $m \in \mathbb{Z} - \{0\}$. Hence the only n which can be invertible in $\mathbb{Z}$ are $n = \pm 1$, and they are invertible, with $n = n^{-1}$. Thus the units of $\mathbb{Z}$ are just ± 1.

Note that in the above example, we used size (absolute value) to help us determine what the units are in $\mathbb{Z}$. We can do something similar for quadratic rings, since we also have a notions of size for them.

Lemma 2.1.2. *Let $d \in \mathbb{Z}$. Then $u \in \mathbb{Z}[\sqrt{d}]$ is a unit if and only if $N(u) = \pm 1$, i.e., if and only if $|N(u)| = 1$.*

Proof. Note we have already done the case of $\mathbb{Z}[\sqrt{d}] = \mathbb{Z}$ in the above example, so we may assume d is a non-square if we wish.

Suppose $u \in \mathbb{Z}[\sqrt{d}]$ is a unit. Then $u^{-1} \in \mathbb{Z}[\sqrt{d}]$ with $uu^{-1} = 1$. Taking the norm of this equation, and using multiplicativity of the norm, we have

$$1 = N(1) = N(uu^{-1}) = N(u)N(u^{-1}).$$

Since $N(u), N(u^{-1}) \in \mathbb{Z}$, this means they are units in $\mathbb{Z}$, and thus ± 1 by the previous example.

Conversely, if $N(u) = u\overline{u} = \pm 1$, then $\pm\overline{u} \in R$ and $u(\pm\overline{u}) = 1$. (Here the $\pm$ sign is the same as in $N(u) = \pm 1$.) □

Hence, thinking of the norm as measuring size in $\mathbb{Z}[\sqrt{d}]$, the units of $\mathbb{Z}[\sqrt{d}]$ are the nonzero elements of $\mathbb{Z}[\sqrt{d}]$ which are as "small" as possible. Moreover, $x \in \mathbb{Z}[\sqrt{d}]$ is a nonzero non-unit if and only if $|N(x)| > 1$.

Exercise 2.1.2. Let $d \in \mathbb{Z}$, and suppose u is a unit in $\mathbb{Z}[\sqrt{d}]$. Show $N(u^{-1}) = N(u)$.

Example 2.1.2. Let $R = \mathbb{Z}[i]$. Then $u = a + bi \in \mathbb{Z}[i]$ can only be a unit if $N(u) = a^2 + b^2 = 1$, i.e., only if $u = \pm 1, \pm i$. Indeed, these are all units as $u^{-1} = u$ if $u = \pm 1$ and $u^{-1} = -u$ if $u = \pm i$. Hence the units of $\mathbb{Z}[i]$ are $\{1, -1, i, -i\}$, i.e., the 4 roots of unity in $\mathbb{Z}[i]$.

[2] For those that have seen some ring theory before: this terminology we will introduce is usually just given for (commutative) rings without zero divisors, which are called *integral domains*. You may assume this if you wish. If you have no idea what I am talking about, just move along. Nothing to see here.

Example 2.1.3. Let $R = \mathbb{Z}[\zeta_n]$. Then each root of unity $u = \zeta_n^j$ is a unit in R: it's inverse is just $u^{-1} = \zeta_n^{n-j} \in R$. Note it is not necessarily just powers of ζ_n that are units in R—e.g., recall we have $\zeta_6 \in \mathbb{Z}[\zeta_3]$ and it is easy to see ζ_6 is a unit in $\mathbb{Z}[\zeta_3]$. However, one can show that the units in $R = \mathbb{Z}[\zeta_n]$ (or in any imaginary quadratic ring R) are precisely roots of unity which are contained in R, which are either the collection of the n-th roots of unity or the $2n$-th roots of unity, depending on whether n is even or odd.

In the above two examples, there are were only finitely many units. We will see in Chapter 5 that real quadratic rings have infinitely many units: they are the integer solutions to $x^2 - dy^2 = \pm 1$ for some $d > 0$. However imaginary quadratic rings (and cyclotomic rings) always have finitely many units, which makes them easier to deal with in some sense. It is easy to determine the units of imaginary quadratic rings:

Proposition 2.1.3. *Let R be $\mathbb{Z}$ or an imaginary quadratic ring $\mathbb{Z}[\sqrt{-d}]$ for some $d > 0$. Then the set of units in R is simply $\{\pm 1\}$ except in the special case $R = \mathbb{Z}[i]$ when it is $\{\pm 1, \pm i\}$.*

Note the above statement needs to be changed if we allow for more general imaginary quadratic rings, like $\mathbb{Z}[\zeta_3]$. But even then, it turns out that $\mathbb{Z}[i]$ and $\mathbb{Z}[\zeta_3]$ are the only imaginary quadratic rings with more than 2 units.

Proof. We've already treated $\mathbb{Z}$ and $\mathbb{Z}[i]$ in examples above, so we just need to consider $R = \mathbb{Z}[\sqrt{-d}]$ for $d > 1$ and show any unit of R must be ± 1. Let $u = a + b\sqrt{-d}$ be such a unit. Then $N(u) = a^2 + db^2 = 1$ by Lemma 2.1.2. (Note for imaginary quadratic fields, since the norm is never negative, we can't have $N(u) = -1$.) But $b \neq 0$ implies $N(u) > 1$. Hence we must have $u = a \in \mathbb{Z}$, whence $u = \pm 1$. □

Definition 2.1.4. *Let R be a ring and $x \in R$ be a nonzero non-unit. We say x is* **reducible** *if there exist nonzero non-units $a, b \in R$ such that $x = ab$. Otherwise, we say x is* **irreducible**.

The condition on x in the first sentence means that will not consider 0 or units to be reducible or irreducible.

Example 2.1.4. Let R be a field. Then any nonzero $u \in R$ is a unit, by the definition of field. Hence we do not consider any elements of a field to be reducible or irreducible. The point is that there's not much sense in talking about factorization in fields, as we can always pull out any factor we want. For instance, think about $\mathbb{Q}$. Take any nonzero $x \in \mathbb{Q}$, say $x = 3$. Given any nonzero $y \in \mathbb{Q}$, e.g., $y = \frac{53}{2}$, we can factor y out of x via $x = y \cdot \frac{x}{y}$, e.g., $3 = \frac{53}{2}\frac{6}{53}$.

Example 2.1.5. Let $R = \mathbb{Z}$. Then $n \in R$ being irreducible just means that $n \notin \{-1, 0, 1\}$ and n cannot be written as a product of two numbers except in trivial ways like $n = 1 \cdot n$ or $n = (-1)(-n)$. Hence n irreducible just means that $n = \pm p$ for some prime $p \in \mathbb{N}$.

Note that the above definition of irreducible is essentially the same as our definition for prime in $\mathbb{N}$ (elements of $\mathbb{N}$ which have exactly 2 factors). Being irreducible in a ring R essentially means we can't break it into "smaller" factors. You might want to call such elements prime, however the word prime is reserved for having a further property which we will define below. (Of course, it turns out for $\mathbb{Z}$, all irreducibles have this property, so being irreducible will be the same as being prime.)

Example 2.1.6. Let $R = \mathbb{Z}[i]$. Then $2 = (1+i)(1-i)$ is a product of two non-units, so 2 is reducible in $\mathbb{Z}[i]$. On the other hand, we can show $1+i$ and $1-i$ are irreducible in $\mathbb{Z}[i]$. Let $x = 1 \pm i$. Then $N(x) = 1^2 + 1^2 = 2$, so x is a nonzero non-unit. If x is reducible, we have $x = ab$ for some nonzero non-units $a, b \in \mathbb{Z}[i]$. Then $2 = N(x) = N(a)N(b)$ and $|N(a)|, |N(b)| > 1$, but the latter condition implies $|N(a)N(b)| \geq 2 \cdot 2 \geq 4$, a contradiction. Thus x must be irreducible.

The argument in the above example generalizes:

Exercise 2.1.3. Let $d \in \mathbb{Z}$ and $x \in \mathbb{Z}[\sqrt{d}]$. Show that if $|N(x)|$ is a prime in $\mathbb{N}$, then x is irreducible.

Exercise 2.1.4. Show 17 is reducible in $\mathbb{Z}[i]$, and find a factorization of 17 into irreducibles.

Exercise 2.1.5. Show 3 is irreducible in $\mathbb{Z}[i]$, even though $N(3)$ is not prime.

The first step in factorization is noting that we can always break (nonzero non-unit) elements up into a product of irreducibles, i.e., we have some factorization.

Proposition 2.1.5 (Existence of factorization)**.** *Let $d \in \mathbb{Z}$. Then any non-zero nonunit $x \in \mathbb{Z}[\sqrt{d}]$ can be factored into irreducibles: $x = a_1 a_2 \cdots a_r$ for some irreducibles $a_1, \ldots, a_r$ in $\mathbb{Z}[\sqrt{d}]$.*

Proof. This proof follows the same descent strategy we employed in the case of $\mathbb{Z}$ in Proposition 1.1.3, so we will be briefer in our explanation of this proof.

Either x itself is irreducible or not. If it is irreducible, then we can take $r = 1$ and $a_1 = x$, and we are done. So assume x is reducible. Then we can write $x = y_1 y_2$ for some nonzero non-units $y_1, y_2 \in \mathbb{Z}[\sqrt{d}]$. By multiplicativity of the norm, we have $N(x) = N(y_1)N(y_2)$ so $|N(x)| = |N(y_1)||N(y_2)|$. Since y_1, y_2 are nonzero non-units, we must have $1 < |N(y_1)|, |N(y_2)| < |N(x)|$.

Now it suffices to show that y_1, y_2 factor into irreducibles. We simply repeat the above argument, which must eventually terminate by descent on the absolute norm. Thus x factors into irreducibles by a similar argument as in Proposition 1.1.3. □

We note there are rings where not all elements are *finite* products of irreducible elements, but may be infinite products of irreducible elements. However, we will not work with such rings in this course.

2.2 Primes and unique factorization

Definition 2.2.1. *Let $a, b \in R$. We say b* **divides** *a, or b is a* **divisor** *of a and write $b \mid a$, if $a = bc$ for some $c \in R$. If b does not divide a, we write $b \nmid a$.*

One way to think about role of units in arithmetic is that multiplication by units does not affect the divisibility of numbers. More precisely:

Exercise 2.2.1. Let $a, a', b \in R$. We say a' is an **associate** of a if $a' = au$ for some unit u of R.

(i) Suppose a' is an associate of a. Show $b \mid a \iff b \mid a'$, i.e., a and a' have precisely the same divisors.

(ii) Suppose $R = \mathbb{Z}[\sqrt{d}]$. Show conversely that if a, a' have exactly the same divisors, then a' is an associate of a.

Definition 2.2.2. *Let $p \in R$ be a nonzero non-unit. If for all $a, b \in R$, $p \mid ab$ implies $p \mid a$ or $p \mid b$, we call p* **prime**. *If every irreducible in R is prime, we say R has the* **prime divisor property**.

That is to say, a prime is something with the property that if it divides the product of two things, it must divide one or the other (and possibly both). It is not true that irreducible elements are always prime (e.g., Example 2.2.2 below), and this issue is intimately tied up with unique factorization. On the other hand, we can prove that prime elements are automatically irreducible.

Proposition 2.2.3. *Let $d \in \mathbb{Z}$. If p is a prime in $\mathbb{Z}[\sqrt{d}]$, then p is irreducible.*

Proof. (Contradiction.) Suppose $p \in \mathbb{Z}[\sqrt{d}]$ is prime, but p is reducible. Say $p = ab$ where a, b are nonzero nonunits. So $|N(a)|, |N(b)| > 1$ and $|N(p)| = |N(a)||N(b)|$ then implies $|N(a)|, |N(b)| < |N(p)|$. Then $p \mid ab$ so $p \mid a$ or $p \mid b$ by primality. Interchanging a and b if necessary, we may assume $p \mid a$. Hence $a = pc$ for some $c \in \mathbb{Z}[\sqrt{d}]$. But then $|N(a)| = |N(p)||N(c)|$ implies $|N(p)| \leq N(a)$, contradicting $|N(a)| < |N(p)|$. □

Next we will show that the prime divisor property is precisely what we need for unique factorization.

Definition 2.2.4. *We say a ring R has* **unique factorization** *if (i) any nonzero non-unit $x \in R$ has a factorization $x = a_1 \cdots a_r$ into irreducibles, and (ii) any two factorizations of $x = a_1 \cdots a_r = b_1 \cdots b_s$ into irreducibles are the same up to ordering and units, i.e., after relabeling b_j's if necessary, we have $s = r$ and there exist units $u_1, \dots, u_r \in R$ such that*

$$b_1 = u_1 a_1, \quad b_2 = u_2 a_2, \quad \dots, \quad b_r = u_r a_r.$$

The first part of the definition just says that we can always factor elements of R (besides 0 and units), which we already know for $R = \mathbb{Z}[\sqrt{d}]$ (Proposition 2.1.5). The second part of the definition is the uniqueness statement. Let's think about what it says for $\mathbb{Z}$. E.g., take $n = -12$. The irreducible factors of n are ± 2 and ± 3, and there are different ways we can write n as a product of irreducibles, e.g.,

$$-12 = (-2) \cdot 2 \cdot 3 = (-3)(-2)(-2).$$

However, these different factorizations are not essentially different: they only differ up to order and signs (units). Reordering the second factorization as $(-2)(-2)(3)$, we see that all the factors match up up to units:

$$-2 = -2, \quad -2 = (-1)2, \quad (-3) = (-1)3.$$

This is what the second part of the definition of unique factorization is about.

Here is another example of two factorizations which are the same up to ordering and units, which is perhaps less obvious at first glance.

Example 2.2.1. In $\mathbb{Z}[i]$ we have the factorizations

$$5 = (2+i)(2-i) = (1+2i)(1-2i).$$

By Exercise 2.1.3, the elements $2+i$, $2-i$, $1+2i$ and $1-2i$ are all irreducible as their absolute norms are all 5. We claim these factorizations are the same up to ordering and units. Recall the units of $\mathbb{Z}[i]$ are $\pm 1, \pm i$. Note $i(2+i) = 2i-1$ and $i(2-i) = 2i+1$. Hence we can write each factor of the second factorization above as a unit times a factor of the first factorization as:

$$1+2i = i(2-i), \quad 1-2i = (-i)(2+i).$$

Exercise 2.2.2. Show that $\mathbb{Z}[\sqrt{-3}]$ has no element of norm 2. Deduce that if $x \in \mathbb{Z}[\sqrt{-3}]$ with $N(x) = 2p$ for some prime p, then x is irreducible.

Example 2.2.2. In $\mathbb{Z}[\sqrt{-3}]$ we have the factorizations

$$4 = 2 \cdot 2 = (1+\sqrt{-3})(1-\sqrt{-3}).$$

Since 2, $1+\sqrt{-3}$, and $1-\sqrt{-3}$ all have norm 4, both of these are irreducible factorizations by the previous exercise. But by Proposition 2.1.3, the only units of $\mathbb{Z}[\sqrt{-3}]$ are ± 1 so $1 \pm \sqrt{-3}$ does not differ from 2 by a unit. Hence these two irreducible factorizations are truly different: they are not the same up to ordering and unit.

Consequently, this failure of unique factorization demonstrates irreducibles which are not prime, i.e., do not satisfy the prime divisor property. Namely $2 \mid (1+\sqrt{-3})(1-\sqrt{-3}) = 4$, but $2 \mid (1 \pm \sqrt{-3})$ because the $1 \pm \sqrt{-3}$ is irreducible and does not differ 2 by a unit. (Alternatively, you can also easily prove this by contradiction by writing $1 \pm \sqrt{-3} = 2(a + b\sqrt{-3})$ and solving for a, b.) Hence 2 is irreducible but not prime in $\mathbb{Z}[\sqrt{-3}]$. A similar argument shows $1 \pm \sqrt{-3}$ is also irreducible but not prime in $\mathbb{Z}[\sqrt{-3}]$.

Theorem 2.2.5. *Let $d \in \mathbb{Z}$. If $\mathbb{Z}[\sqrt{d}]$ has the prime divisor property, then $\mathbb{Z}[\sqrt{d}]$ has unique factorization.*

Proof. By Proposition 2.1.5, we already know the existence of irreducible factorizations, so it suffices to check the second part of the definition of unique factorization. We do this

by contradiction. Namely, assume $\mathbb{Z}[\sqrt{d}]$ has the prime divisor property, but there exists a non-zero nonunit x which has two irreducible factorizations

$$x = a_1 \cdots a_r = b_1 \cdots b_s$$

that are not the same up to ordering and units.

If some factors of these two factorization are the same up to units, e.g., $a_r = ub_s$ so multiplying by a_r^{-1} gives $a_1 \cdots a_{r-1} = ub_1 \cdots b_{s-1}$, we can cancel off any common (up to units) factors, and reorder/relabel to get two nonempty collections of irreducibles $a_1, \dots a_m$ and $b_1, \dots, b_n$ and a unit u such that

$$a_1 \cdots a_m = ub_1 \cdots b_n \tag{2.2.1}$$

but no a_i differs (multiplicatively) from any b_j by a unit. Put another way, no a_i divides any b_j.

Necessarily $m, n > 1$. To see this, first note that $m = n = 1$ implies $a_1 = ub_1$ so a_1 and b_1 differ units, which would be a contradiction. Hence at least one of m and n is bigger than 1, say $n > 1$ but $m = 1$. Then $a_1 = ub_1 \cdots b_n$, contradicting the irreducibility of a_1. Similarly $m > 1$ and $n = 1$ is impossible, so $m, n > 1$ as claimed. (Actually, we only need $n > 1$ below.)

Then (2.2.1) implies

$$a_1 \mid b_1 \cdots b_n = b_1 \cdot (b_2 \cdots b_n),$$

so by the prime divisor property $a_1 \mid b_1$ or $a_1 \mid b_2 \cdots b_n$. The former is impossible as we have already canceled common (up to units) factors, so we must have $a_1 \mid b_2 \cdots b_n$. If $n = 2$, this means $a_1 \mid b_2$, which is again impossible. Hence $n > 2$ and we have

$$a_1 \mid b_2(b_3 \cdots b_n).$$

Repeating this argument with the prime divisor property (i.e., use descent), we see that in the end we have $a_1 \mid b_n$, giving us our desired contradiction. □

Exercise 2.2.3. Let $R = \mathbb{Z}[\sqrt{d}]$ for some $d \in \mathbb{Z}$. Deduce from the above theorem that R has unique factorization if and only if it has the prime divisor property. (*Suggestion:* Reread Example 2.2.2.)

We remark that the results in this section and the previous one apply to more general rings than those of the form $R = \mathbb{Z}[\sqrt{d}]$. The key feature we needed in all of the proofs was the existence of a norm map $N : R \to \mathbb{Z}$ with the properties listed at the beginning of this section. Cyclotomic rings, and many other rings, have such a norm map. On the other hand, rings like $\mathbb{Z}/n\mathbb{Z}$ do not. However, we will see later that most nonzero elements of $\mathbb{Z}/n\mathbb{Z}$ are units, so there is not much point in talking about factorization in $\mathbb{Z}/n\mathbb{Z}$ anyway.

2.3 The Euclidean algorithm

In the last section, we showed that unique factorization follows from the prime divisor property for quadratic rings $\mathbb{Z}[\sqrt{d}]$. (In fact they are equivalent by Exercise 2.2.3.) Now you are probably wondering, okay, so how do we prove the prime divisor property for various rings like $\mathbb{Z}$ and $\mathbb{Z}[i]$? The simplest method that I know for $\mathbb{Z}$ and $\mathbb{Z}[i]$ (and a few other rings) is via the Euclidean algorithm, which is a way to compute gcds. While this method does not work for all quadratic rings which happen to have unique factorization, it will be enough to prove unique factorization in all cases we will use in this course. (Both some rings with unique factorization, as well as all quadratic rings without unique factorization will not possess a Euclidean algorithm.) In this section, we'll review the Euclidean algorithm in the classical case of integers, and see how it yields the prime divisor property, and thus the fundamental theorem of arithmetic. Then we'll look at the Euclidean algorithm for quadratic fields $\mathbb{Z}[\sqrt{d}]$ in the next section.

Recall for natural numbers a, b, their **gcd** or **greatest common divisor**, denoted $\gcd(a, b)$ is the largest natural number d such that $d \mid a$ and $d \mid b$. Note that $\gcd(a, b)$ exists for all $a, b \in \mathbb{N}$ by the fact that any divisor of a is at most a and descent. There are different versions and variations of the classical Euclidean algorithm. We present three. While only the first is needed to prove the prime divisor property for $\mathbb{Z}$, the second is useful for extending the Euclidean algorithm to $\mathbb{Z}[i]$. The third will be a variant that we use for a quick detour for describing how to solve another basic number theory problem: how to solve linear Diophantine equations (in 2 variables).

The gcd by subtraction

Let $a, b, d \in \mathbb{N}$.

First note that if d is a **common divisor** of a and b, i.e.,

$$a = a'd, \quad b = b'd,$$

for some $a', b' \in \mathbb{N}$, then

$$a - b = a'd - b'd = (a' - b')d$$

so d is a divisor of $a - b$. Similarly, if d is a common divisor of $a - b$ and b, then it is also a divisor of $a = (a - b) - b$. Hence the common divisors of a and b are the same as the common divisors of $a - b$ and b. In particular,

$$\gcd(a, b) = \gcd(b, a - b)$$

Euclid used this idea to make an efficient algorithm to determine $\gcd(a, b)$.

The **Euclidean algorithm** goes as follows. Set

$$a_1 = \max\{a, b\}, \quad b_1 = \min\{a, b\}.$$

Then we inductively compute

$$a_{i+1} = \max\{b_i, a_i - b_i\}, \quad b_{i+1} = \min\{b_i, a_i - b_i\},$$

stopping only when we have

$$a_k = b_k.$$

This procedure produces smaller and smaller pairs of natural numbers so must eventually terminate by descent.[3] The max/min business is to ensure we always have $a_i \geq b_i$ so that the $a_i - b_i$ appearing in the next step is positive. You might find it easier to think of this algorithm as defining a_{i+1}, b_{i+1} so that

$$\{a_{i+1}, b_{i+1}\} = \{b_i, a_i - b_i\} \quad \text{and} \quad a_{i+1} \geq b_{i+1}.$$

The reason this works is as follows. Since $\gcd(a,b) = \gcd(b, a-b)$, we have

$$\gcd(a,b) = \gcd(a_1, b_1) = \gcd(a_2, b_2) = \cdots = \gcd(a_k, b_k) = \gcd(a_k, a_k) = a_k.$$

Example 2.3.1. Let $a_1 = a = 15$, $b_1 = b = 6$. Then $\{b_1, a_1 - b_1\} = \{6, 9\}$, so we set $a_2 = 9$, $b_2 = 6$. Then $\{b_2, a_2 - b_2\} = \{6, 3\}$, so we set $a_3 = 6$, $b_3 = 3$. Similarly, we get $a_4 = 3$, $b_4 = 3$, at which point the algorithm terminates leaving us with $\gcd(15, 6) = 3$. Alternatively, without explicitly writing out all the a_i's and b_i's, we can write the Euclidean algorithm as

$$\gcd(15,6) = \gcd(9,6) = \gcd(6,3) = \gcd(3,3) = 3.$$

Example 2.3.2. Consider $a = 18$, $b = 5$. Then we have

$$\gcd(18,5) = \gcd(13,5) = \gcd(8,5) = \gcd(5,3) = \gcd(3,2) = \gcd(2,1) = \gcd(1,1) = 1.$$

If $\gcd(a,b) = 1$, we say a and b are **coprime** or **relatively prime**.

Exercise 2.3.1. Compute $\gcd(84, 63)$ using the above method. Write out each step.

The gcd by division with remainder

A more efficient version of the Euclidean algorithm is as follows. Given $a, b \in \mathbb{N}$, with $a \geq b$, we can write $a = qb + r$ for unique $q \in \mathbb{N}$, $r \in \mathbb{Z}_{\geq 0}$. We call r the **remainder** of a/b. Set

$$a_1 = \max\{a,b\}, \quad b_1 = \min\{a,b\},$$

$$a_{i+1} = b_i, \quad b_{i+1} = \text{remainder of } a_i/b_i,$$

halting when we have a pair

$$(a_k, b_k) \text{ with } b_k \mid a_k.$$

Then

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

This algorithm is essentially the same as the subtraction version, but the division can do several steps of subtraction at once.

[3] Keep this in mind for generalization to quadratic rings.

Example 2.3.3. Let's revisit Example 2.3.1, i.e., consider $a_1 = a = 18$, $b_1 = b = 5$. We have $18 = 3 \cdot 5 + 3$, so we set $a_2 = 5$ and $b_2 = 3$ (this 3 is the second one, i.e., the remainder, not the one in front of the 5). Then we write $5 = 1 \cdot 3 + 2$, so we set $a_3 = 3$, $b_3 = 2$. Similarly, we'll get $a_4 = 2$, $b_4 = 1$, at which point we conclude the gcd is $b_4 = 1$. Writing things without the a_i's and b_i's would be

$$\gcd(18, 5) = \gcd(5, 3) = \gcd(3, 2) = \gcd(2, 1) = 1,$$

so we've saved a couple of steps from the subtraction version, at the expense of needing to do intermediate division calculations.

Write a and b in binary. Suppose $a > b$ and a is n bits (binary digits) long. Then the remainder in a/b has at most $n-1$ bits, so this algorithm will terminate at most n steps. In other words, if $\max a, b < 2^{n+1}$, then we can determine $\gcd(a, b)$ in at most n steps. This is as efficient as one could hope for. A computer can handle numbers thousands of digits long in seconds. We note that computing gcd's is much easier than factoring numbers—in particular determining if 2 numbers are coprime is much easier than determining if a given number is prime. While you can compute gcd's by factoring and looking at the prime (power) factors in common, this is very inefficient for large numbers, and the Euclidean algorithm (either version) is much much better. For instance, even with very advanced algorithms, a modern computer might take up a year to factor a 200-digit number. This latter fact is important in modern cryptography, as we will discuss in the next chapter.

Another advantage of the division version is it can deal with other number systems. E.g., if you want to compute $\gcd(17, 4 + i)$ in $\mathbb{Z}[i]$, you can divide 17 by $4 + i$ (and get $4 - i$ exactly), but subtraction gives you nothing.

Exercise 2.3.2. Compute $\gcd(42, 8)$ using the division method. Write out each step.

Linear Diophantine equations

If we go back to the subtraction version of the Euclidean algorithm, it is clear that at each step a_i and b_i are (integral) linear combinations of a and b. Hence

$$\gcd(a, b) = a_k = ma + nb \tag{2.3.1}$$

for some $m, n \in \mathbb{Z}$.

Recall number theory is about determining integer solutions to Diophantine equations. The simplest kind of Diophantine equations are linear ones, and the Euclidean algorithm tells us about such equations in the simplest (nontrivial) case of two variables, i.e., equations of the form

$$ax + by = c,$$

for $a, b, c \in \mathbb{Z}$. Here we will apply the Euclidean algorithm to determine completely when this equation has an integer solution (i.e., a solution $(x, y) \in \mathbb{Z} \times \mathbb{Z}$), and when it does, how to find all of them.

Note if one of a, b is zero, this degenerates to a trivial situation (e.g., $ax = c$ or $0 = c$), so let's assume $a, b \in \mathbb{Z}$ are nonzero. We extend the **gcd** to nonzero $a, b \in \mathbb{Z}$ by setting $\gcd(a, b) = \gcd(|a|, |b|)$. (We will give a more general definition of gcd later.)

Proposition 2.3.1. *Let $a, b \in \mathbb{Z}$ be nonzero. Then $ax + by = c$ has an integer solution $(x, y) \in \mathbb{Z} \times \mathbb{Z}$ if and only if $\gcd(a, b) \mid c$.*

Proof. ($\Rightarrow$) If there is a solution, then

$$\gcd(a, b) \mid ax \text{ and } \gcd(a, b) \mid by \implies \gcd(a, b) \mid c.$$

($\Leftarrow$) If $\gcd(a, b) \mid c$, we can write $c = \gcd(a, b)d$. By the Euclidean algorithm we have $\gcd(a, b) = am + bn$ as in (2.3.1), for for some $m, n \in \mathbb{Z}$, which implies

$$c = \gcd(a, b)d = amd + bnd.$$

□

To actually find solutions to an equation $ax+by = c$, we need not only $\gcd(a, b)$ but also the m and n in (2.3.1). This can be done through a variety of equivalent methods, sometimes called the **extended Euclidean algorithm**. We will present the **tableau method**, which is more efficient than version many number theory texts present. For simplicity, we just present this method by way of example.

Consider $a = 34$, $b = 19$. The idea is to use a little linear algebra, and is similar to matrix row reduction, but we build a table, starting with the following two rows. For clarification I will write the underlying equation on the right, though in practice you will omit this.

m	n	x	$\longleftrightarrow$	$ma + nb = x$
1	0	34		$1 \cdot a + 0 \cdot a = 34$
0	1	19		$0 \cdot a + 1 \cdot b = 19$

The entries running down the x column will just be the successive numbers $a_1, b_1, b_2, \ldots, b_k$ from the division algorithm. The m and n entries for the b_i row will just be the coefficients needed for $ma + nb = b_i$. For example, here $b_2 = a_1 - b_1$, so the next row will just be obtained by subtracting the second from the first (do this to each column) to get

m	n	x	$\longleftrightarrow$	$ma + nb = x$
1	0	34		$1 \cdot a + 0 \cdot a = 34$
0	1	19		$0 \cdot a + 1 \cdot b = 19$
1	1−	15		$1 \cdot a - 1 \cdot b = 15$

We do this again to get

m	n	x	$\longleftrightarrow$	$ma + nb = x$
1	0	34		$1 \cdot a + 0 \cdot a = 34$
0	1	19		$0 \cdot a + 1 \cdot b = 19$
1	-1	15		$1 \cdot a - 1 \cdot b = 15$
-1	2	4		$-1 \cdot a + 2 \cdot b = 4$

Now 4 goes into 15 3 times, so we should subtract 3 times the last row from the previous row to get

m	n	x	$\longleftrightarrow$	$ma + nb = x$
1	0	34		$1 \cdot a + 0 \cdot a = 34$
0	1	19		$0 \cdot a + 1 \cdot b = 19$
1	-1	15		$1 \cdot a - 1 \cdot b = 15$
-1	2	4		$-1 \cdot a + 2 \cdot b = 4$
4	-7	3		$4 \cdot a - 7 \cdot b = 3$

With one more step we are done:

m	n	x	$\longleftrightarrow$	$ma + nb = x$
1	0	34		$1 \cdot a + 0 \cdot a = 34$
0	1	19		$0 \cdot a + 1 \cdot b = 19$
1	-1	15		$1 \cdot a - 1 \cdot b = 15$
-1	2	4		$-1 \cdot a + 2 \cdot b = 4$
4	-7	3		$4 \cdot a - 7 \cdot b = 3$
-5	9	1		$-5 \cdot a + 9 \cdot b = 1$

We know we are done now because the last b_j ($x = 1$) divides the previous b_j ($x = 3$). Hence the tableau method has shown two things:

$$\gcd(a, b) = \gcd(34, 19) = 1,$$

which one already gets from the usual Euclidean algorithm, and

$$\gcd(a, b) = -5a + 9b, \quad \text{i.e.} \quad 34(-5) + 19(9) = 1,$$

which one does not. From the latter fact, we can explicitly find integer solutions x, y to

$$ax + by = 34x + 19y = c,$$

for any $c \in \mathbb{Z}$, as in the proof of Proposition 2.3.1. For instance, if $c = 5$, then we can multiply the equation before the last by 5 to get a solution

$$34(-25) + 19(45) = 5,$$

i.e., $x = 5(-5)$, $y = 9(5)$ is a solution to $34x + 19y = 5$.

Exercise 2.3.3. Use the tableau method to compute $\gcd(120, 39)$, and use the outcome to find an integer solution to $120x + 39y = 6$.

Now that we know precisely when a 2-variable linear Diophantine equation $ax + by = c$ is solvable, and how to find a single solution, you might ask how do we determine all integer solutions. Because the equation is linear, we can do the same thing one does in linear algebra: we can combine one inhomogeneous solution (the case with $c \neq 0$) with all homogenous solutions (the case with $c = 0$). There are infinitely many homogenous solutions, and they are easy to describe:

Exercise 2.3.4. Let $a, b \in \mathbb{Z}$ be nonzero, and put $a' = a/\gcd(a,b)$, $b' = b/\gcd(a,b)$. Show that the set of solutions to the homogeneous equation

$$ax + by = 0 \tag{2.3.2}$$

are precisely the set

$$\{(x,y) = (kb', -ka') : k \in \mathbb{Z}\}.$$

Proposition 2.3.2. *Let $a, b, c \in \mathbb{Z}$ with a, b nonzero. Suppose $ax + by = c$ has a solution $(x_0, y_0) \in \mathbb{Z} \times \mathbb{Z}$. Then the set of integer solutions to $ax + by = c$ are the ordered pairs of the form $(x_0, y_0) + (x, y)$ where (x, y) is a solution to the homogenous equation (2.3.2).*

Proof. You should have seen this proof in linear algebra already.

($\Rightarrow$) Suppose (x_1, y_1) is another solution to $ax + by = c$. We want to show it is of the desired form. Then

$$(ax_1 + by_1) - (ax_0 + by_0) = c - c = 0.$$

Hence $(x, y) = (x_1 - x_0, y_1 - y_0)$ is a solution to (2.3.2).

($\Leftarrow$) Suppose (x, y) is a solution to (2.3.2). Then

$$a(x_0 + x) + b(y_0 + y) = (ax_0 + by_0) + (ax + by) = c + 0 = c$$

so $(x_0, y_0) + (x, y)$ is a solution to $ax + by = c$. □

Exercise 2.3.5. Find all integer solutions to $12x + 35y = 3$.

Unique factorization for $\mathbb{Z}$

Here we will finally complete the proof of the fundamental theorem of arithmetic (stated variously as Theorem 1.1.1, Theorem 2.0.1 and Theorem 2.0.2). By Theorem 2.2.5, it suffices to prove $\mathbb{Z}$ has the prime divisor property:

Theorem 2.3.3 (Prime divisor property)**.** *Let $p \in \mathbb{Z}$ be irreducible, and $a, b \in \mathbb{Z}$ with $a, b \notin \{0, 1, -1\}$. If $p \mid ab$, then $p \mid a$ or $p \mid b$. In other words, every irreducible in $\mathbb{Z}$ is prime.*

Remark 2.3.4. A consequence of our way of defining primes in rings means that negative numbers in $\mathbb{Z}$ are also prime, i.e., the primes of $\mathbb{Z}$ are $\pm 2, \pm 3, \pm 5, \pm 7, \pm 11, \ldots$. This may seem strange, but the reason is made more clear when thinking about how to generalize the notion of prime to other rings like $\mathbb{Z}[i]$, where one doesn't have the notion of positive versus negative. The point is in general there is no natural way to distinguish one irreducible p from the set of multiples up where u ranges over units, so it is easiest to call each up prime if p is. Of course, when we're working with just the usual integers, it typically suffices to restrict to positive primes, so in the future when we say something like "Let p be a prime (number)" without other qualification/context, we will mean p is a prime in $\mathbb{N}$, i.e., a positive prime in $\mathbb{Z}$ by default. If we mean p can be negative, we will say something like "Let $p \in \mathbb{Z}$ be prime."

Proof. Suppose $p \mid ab$, but suppose $p \nmid a$. Since p is irreducible but $p \nmid a$, the only possible common divisors of p and a are units, i.e., ± 1, so $\gcd(p, a) = 1$. Then by Proposition 2.3.1, there exist $m, n \in \mathbb{Z}$ such that

$$am + pn = 1.$$

Multiplying by b gives

$$abm + pbn = b.$$

Now $p \mid ab$ by assumption, and clearly $p \mid pbn$, so it divides the left hand side of this equation, and therefore the right, i.e., $p \mid b$, which is what we wanted to prove. □

This completes the fundamental theorem of arithmetic, i.e., uniqueness of prime factorization for $\mathbb{Z}$.

Exercise 2.3.6. Recall the proof of Lemma 1.5.2, which required using prime factorization, i.e., the fundamental theorem of arithmetic. Does the proof only need the existence of factorization or does it require uniqueness as well? Explain.

2.4 A Euclidean algorithm for (two) quadratic rings

What makes the Euclidean algorithm in $\mathbb{Z}$ work was the fact that we could write $\gcd(a_i, b_i) = \gcd(a_{i+1}, b_{i+1})$ where a_{i+1} and b_{i+1} are smaller than a_i and b_i. This was used in both the subtraction and division versions of the Euclidean algorithm.

For quadratic rings $\mathbb{Z}[\sqrt{d}]$, recall the notion of size is given by the norm or absolute norm. However, it is not true in general that if b is "smaller" than a, then $a - b$ is "smaller" than a in $\mathbb{Z}[\sqrt{d}]$.

Example 2.4.1. Consider $a = 1 + 2i$, $b = 1 - i \in \mathbb{Z}[i]$. Then $N(a) = 5$, $N(b) = 2$ but $a - b = 3i$ has norm $N(a - b) = 9$.

This suggests we can't generalize the subtraction version of the Euclidean algorithm to $\mathbb{Z}[i]$ or other quadratic rings, but what about the division version? For that, we used the fact that for $a, b \in \mathbb{Z}$ with $b \neq 0$, we can write

$$a = qb + r, \quad \text{for some } q \in \mathbb{Z},\ |r| < |b|. \tag{2.4.1}$$

(Here I stated this for $\mathbb{Z}$, rather than $\mathbb{N}$, to suggest how to make such a statement for $\mathbb{Z}[\sqrt{d}]$. Before, we also assumed $a \geq b$, but this is not necessary as if $|a| < |b|$ one can take $q = 0$, $r = a$.) In fact, if we assume $r \geq 0$, then q is uniquely determined, but positivity of r is not actually important for the Euclidean algorithm (nor does it make sense for quadratic fields). As before, we can treat quadratic rings together with $\mathbb{Z}$ uniformly.

Definition 2.4.1. *Let $d \in \mathbb{Z}$. We say $\mathbb{Z}[\sqrt{d}]$ has the* **division property** *if for all $a, b \in \mathbb{Z}[\sqrt{d}]$ with $b \neq 0$, there exist $q, r \in \mathbb{Z}[\sqrt{d}]$ such that*

$$a = qb + r, \quad |N(r)| < |N(b)|. \tag{2.4.2}$$

Note that when $\mathbb{Z}[\sqrt{d}] = \mathbb{Z}$, the division property just means (2.4.1), as $|N(r)| = r^2 < |N(b)| = b^2$ is equivalent to $|r| < |b|$. Also note when the division property holds, the q and r in (2.4.2) need not unique, even when $\mathbb{Z}[d] = \mathbb{Z}$, e.g., both $7 = 2 \cdot 3 + 1$ and $7 = 3 \cdot 3 + (-2)$ are in the form (2.4.2).

If $\mathbb{Z}[\sqrt{d}]$ has the division property, then we will get a Euclidean algorithm, and then unique factorization.

Lemma 2.4.2. *Let $a, m \in \mathbb{Z}[\sqrt{d}]$. Then $m \mid a$ implies $N(m) \mid N(a)$ in $\mathbb{Z}$. Moreover, if $d < 0$, then any nonzero $a \in \mathbb{Z}[\sqrt{d}]$ has only finitely many divisors m.*

We'll see later that real quadratic rings have infinitely many units, which implies that numbers in real quadratic rings have infinitely many divisors.

Proof. Write $a = km$ for some $k \in \mathbb{Z}$. Then $N(a) = N(k)N(m)$ by Exercise 1.5.8, hence $N(m) \mid N(a)$ in $\mathbb{Z}$.

Now suppose $d < 0$, which means the norm map is non-negative. Then if $m = x + y\sqrt{d}$ is a divisor of a, we have $N(m) = x^2 - dy^2 \leq x^2 + y^2$. There are only finitely many choices for $(x, y) \in \mathbb{Z} \times \mathbb{Z}$ such that $x^2 + y^2 \leq N(a)$ (e.g., we need $|x|, |y| \leq \sqrt{N(a)}$), and thus only finitely many possibilities for a divisor m of a. □

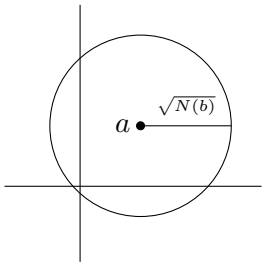
Definition 2.4.3. *Let $a, b \in \mathbb{Z}[\sqrt{d}]$, not both zero. We say d is a* **gcd** *(***greatest common divisor***) of a and b if $|N(d)|$ is maximal among elements such that $d|a$ and $d|b$. Denote the set of all gcds of a and b by* $\mathrm{GCD}(a, b)$.

Note if $m \mid a$ and $m \mid b$, then $N(m) \mid N(a)$ and $N(m) \mid N(b)$, so if at least one of a, b is nonzero, $|N(m)|$ is bounded and the set of all common divisors (which is nonempty as it always contains 1) has an element of maximal absolute norm. Thus $\mathrm{GCD}(a, b)$ always exists. Also, 0 is never a gcd (since 1 is a common divisor), which means if m is a gcd of a and b, so is $-m$, and thus gcds are not unique. More generally, if m is a gcd of a and b, then so is um for any unit u.

For the rest of this section, consider the imaginary quadratic ring $\mathbb{Z}[\sqrt{-d}]$ with $d > 0$. This situation is nicer than the real quadratic case because the norm is always non-negative (it is just the square of a length) and geometrically $\mathbb{Z}[\sqrt{-d}]$ is a lattice in $\mathbb{C}$. In addition, the main applications we have in mind for unique factorization, besides for $\mathbb{Z}$, are for certain imaginary quadratic rings.

First we want to show that there is an algorithm to determine, for given $a, b \in \mathbb{Z}[\sqrt{-d}]$, if there exist q, r satisfying (2.4.2). In other words, does there exist $q \in \mathbb{Z}[\sqrt{-d}]$ such that $r = a - qb$ has norm less than $N(b)$? We may as well assume $N(a) \geq N(b)$, otherwise we just take $q = 0$, $r = a$.

Recall that $N(r) = r\overline{r}$ is simply the square of the distance of r from the origin in $\mathbb{C}$, i.e., the square of the distance of a from qb in $\mathbb{C}$. Hence there exist q, r as in (2.4.2) if and only if there exists $q \in \mathbb{Z}[\sqrt{-d}]$ such that qb lies in the open disc of radius $\sqrt{N(b)}$ about a.



Now the following norm inequality will be useful.

Lemma 2.4.4. *For $a, b \in \mathbb{Z}[\sqrt{-d}]$, we have $N(a+b) \leq N(a) + 2\sqrt{N(a)N(b)} + N(b)$.*

Proof. One can give an algebraic proof, but a geometric one is easier. Let $\ell_1 = \sqrt{N(a)}$ and $\ell_2 = \sqrt{N(b)}$ be the lengths of a and b, thought of as vectors in $\mathbb{C}$. Now the length of $a+b$ is at most $\ell_1 + \ell_2$, from the usual triangle inequality (draw a picture), so $\sqrt{N(a+b)} \leq \ell_1 + \ell_2$, i.e.,

$$N(a+b) \leq (\ell_1 + \ell_2)^2 = N(a) + 2\sqrt{N(a)N(b)} + N(b).$$

□

We remark the bound in the above lemma is attained if (and only if) a and b are vectors in the same direction, i.e., $b = \lambda a$ for some positive $\lambda \in \mathbb{Q}$.

Going back to our problem, if $N(r) = N(a - qb) < N(b)$, then by the lemma (and the assumption $N(a) \geq N(b)$) this means that

$$N(qb) = N(a-r) \leq N(a) + 2\sqrt{N(a)N(r)} + N(r) < N(a) + 2\sqrt{N(a)N(b)} + N(b).$$

This means we just need to consider q with norms $< N(a)/N(b) + \sqrt{N(a)/N(b)} + 1$. Summing up, this gives the following (non-optimized) algorithm.

Division algorithm for $\mathbb{Z}[\sqrt{-d}]$

Problem: Given a, b in $\mathbb{Z}[\sqrt{-d}]$, $b \neq 0$, find $q, r \in \mathbb{Z}[\sqrt{-d}]$ satisfying (2.4.2), i.e., such that $a = qb + r$ and $N(r) < N(b)$—or show no such q, r exist.

(1) If $N(a) < N(b)$, take $q = 0$, $r = a$ and we're done.

(2) Otherwise, determine all $q = x + yi\sqrt{d}$, $x, y \in \mathbb{Z}$ with $N(q) = x^2 + dy^2 < B$, where $B = N(a)/N(b) + \sqrt{N(a)/N(b)} + 1$. (So we only need to check $|x| \leq B$, $|y| \leq B/\sqrt{d}$.)

(3) For all such q, compute $N(a - qb)$. If $N(a - qb) < N(b)$, we may take this q with $r = a - qb$ to satisfy (2.4.2).

(4) If we found no solutions q, r in the previous steps, then there do not exist any, i.e., the division property fails.

We remark that if desired, all solutions to (2.4.2) can be found if desired in step (3). Also, we may consider a larger set of q than given by the bound in (2) if we want. E.g., for implementation on a computer, we could just write code that does Step (3) for all $q = x + yi\sqrt{d}$ with $|x|, |y| \leq B$. This simplifies the programming slightly at the expense of making the computer do slightly more work in (3).

Example 2.4.2. Consider $a = 2 - 3i$, $b = 1 + 2i$ in $\mathbb{Z}[i]$. Note $N(a) = 13$, $N(b) = 5$. We want to consider q with norm $< 13/5 + \sqrt{13/5} + 1 < 6$. It is easy to see this is the set of q of the form $x + yi$ with $|x|, |y| \leq 2$ and $|x|$, $|y|$ not both 2. (There are 21 such q.) In this particular case, we find (by computer) that there are 3 solutions to (2.4.2):

$$q = -1 - i, \quad r = 1, \quad N(r) = 1,$$
$$q = -1 - 2i, \quad r = i - 1, \quad N(r) = 2,$$
$$q = -i, \quad r = -2i, \quad N(r) = 4.$$

Note that not only are q, r are not unique, the norm of the remainders r are not even uniquely determined.

Exercise 2.4.1. Consider $a = 3 - \sqrt{-2}$, $b = 1 + \sqrt{-2}$ in $\mathbb{Z}[\sqrt{-2}]$. Find all $q, r \in \mathbb{Z}[\sqrt{-2}]$ such that $a = qb + r$ with $N(r) < N(b)$.

The next exercise shows $\mathbb{Z}[\sqrt{-3}]$ does not have the division property.

Exercise 2.4.2. Consider $a = 1 + \sqrt{-3}$, $b = 2$ in $\mathbb{Z}[\sqrt{-3}]$. Show there are no $q, r \in \mathbb{Z}[\sqrt{-3}]$ with $a = qb + r$ and $N(r) < N(b)$. (You can use the division algorithm but you don't have to.)

Theorem 2.4.5. *The rings $\mathbb{Z}[i]$ and $\mathbb{Z}[\sqrt{-2}]$ have the division property.*

Proof. Consider the ring $\mathbb{Z}[\sqrt{-d}]$ for $d > 0$. Let $a, b \in \mathbb{Z}[\sqrt{-d}]$ with $b \neq 0$. Any multiple qb of b in $\mathbb{Z}[\sqrt{-d}]$ is over the form $mb + ni\sqrt{d}b$ for some $m, n \in \mathbb{Z}$, i.e., the set of multiples qb of b are the lattice in $\mathbb{C}$ generated by b and $i\sqrt{d}b$. Recall multiplication by i acts as 90-degree rotation about the origin in $\mathbb{C}$, so $i\sqrt{d}b$ the point obtained rotating b 90-degrees and scaling by $\sqrt{d}$.

Now we use the lattice to tile $\mathbb{C}$ by rectangles whose vertices are lattice points qb, specifically translates of the rectangle (by mb and $ni\sqrt{d}b$ for $m, n \in \mathbb{Z}$) with vertices 0, b, $i\sqrt{d}b$ and $b + i\sqrt{d}b$. (See Fig. 2.4.1 for a picture when $d = 1$, $b = 2 + i$.) Each of these rectangles have side lengths $\sqrt{N(b)}$ and $\sqrt{dN(b)}$.

Now a lies in one of these rectangles R. Furthermore some vertex $v = qb$ of R lies within distance δ of a, where δ is one half of the diagonal length of R (the farthest an interior point of R can be from all vertices happens for the midpoint). It is easy to see $\delta = \frac{\sqrt{(1+d)N(b)}}{2}$. If $d \leq 2$, then $\delta < N(b)$. Consequently, $r = a - qb$ satisfies $N(r) < N(b)$. □

We wrote the argument above so you can see that it really only works for $d = 1, 2$. Moreover, if you think about the geometry of the argument, it seems like the division property should fail for $d \geq 3$. Indeed, we saw in Exercise 2.4.2 it fails for $d = 3$. However to prove this more generally, one needs to exhibit a rectangle R as in the above proof and an element of $\mathbb{Z}[\sqrt{-d}]$ (not just $\mathbb{C}$) which is farther away from every vertex of R than the shortest side length of R. I will simply leave the $d = 5$ case for you:

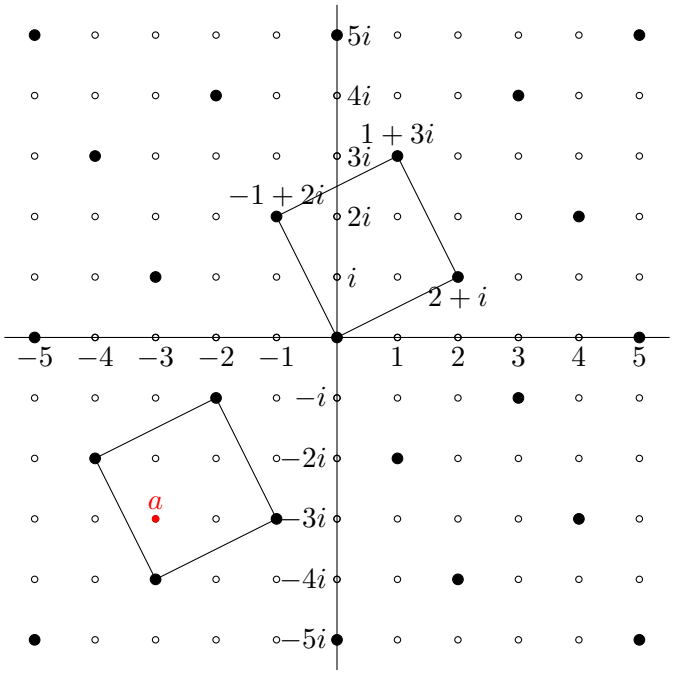


Figure 2.4.1: Multiples of $2+i \in \mathbb{Z}[i]$

Exercise 2.4.3. Show the division property does not hold in $\mathbb{Z}[\sqrt{-5}]$.

Now that we've established the division property for $\mathbb{Z}[i]$ and $\mathbb{Z}[\sqrt{-2}]$, we will see how the Euclidean algorithm extends to these cases.

Euclidean algorithm for $\mathbb{Z}[\sqrt{-d}]$ via division

Problem: Assume $\mathbb{Z}[\sqrt{-d}]$ satisfies the division property, e.g., $d = 1, 2$. Given a, b in $\mathbb{Z}[\sqrt{-d}]$ not both 0, find a gcd m of a and b.

(1) Let $\{a_1, b_1\} = \{a, b\}$ such that $N(a_1) \geq N(b_1)$. Put $i = 1$.

(2) If $b_i = 0$, we can take $m = a_i$ to be a gcd of a and b.

(3) Otherwise, apply the division algorithm to write $a_i = q_i b_i + r_i$ for some $q_i, r_i \in \mathbb{Z}[\sqrt{-d}]$ with $N(r_i) < N(b_i)$.

(4) Let $a_{i+1} = b_i$, $b_{i+1} = r_i$.

(5) Replace i with $i + 1$ and repeat from Step (2).

The proof that this works is similar to the case for $\mathbb{Z}$. With notation as in Step (3), note any divisor of both a_i and b_i is also a divisor of r_i, and conversly any divisor of r_i and b_i is a divisor of a_i, thus $\mathrm{GCD}(a_i, b_i) = \mathrm{GCD}(b_i, r_i) = \mathrm{GCD}(a_{i+1}, b_{i+1})$. Then by descent on

the norm of b_i, eventually some b_k is 0, whence we are led to $\mathrm{GCD}(a,b) = \mathrm{GCD}(a_k, 0)$ Now note that a_k is a gcd of a_k and 0 for any $a_k \neq 0$ so at some point we get a gcd $m = a_k$ in Step (2).

Alternatively, we can replace Step (2) by

(2′) If $b_i \mid a_i$, we can take $m = b_i$ to be a gcd of a and b.

This is true because we get $b_k = 0$ exactly when $b_{k-1} \mid a_{k-1}$. If $b_{k-1} \mid a_{k-1}$, both versions of the algorithm will output $a_k = b_{k-1}$ as a gcd, but the version using (2′) will just make one less pass. and it will simply terminate the algorithm earlier. However I originally wrote the algorithm with the formulation in Step (2) because (i) it is easier to formula descent terminating with $b_i = 0$ rather than the condition $b_i \mid a_i$, and (ii) to test for $b_i \mid a_i$ in general, you need to first apply the division algorithm in Step (3). That said, in some cases it will be obvious that $b_i \mid a_i$ (e.g., if $b_i = 1$ so 1 is a gcd) so when working out examples by hand we may use Step (2′).

Example 2.4.3. Consider $a = 2 - 3i$, $b = 1 + 2i$ in $\mathbb{Z}[i]$ as in Example 2.4.2. Since $N(a) = 13 > N(b) = 5$, we take $a_1 = a$, $b_1 = b$. By Example 2.4.2, there are 3 ways to write $a_1 = q_1 b_1 + r_1$. Let's see how the algorithm proceeds with these different choices.

If we take $q_1 = -1 - i$ then $r_1 = 1$, so we take $a_2 = b_1 = 1 + 2i$ and $b_2 = r_1 = 1$. Since $b_2 = 1$, Step (2') tells us 1 is a gcd of a and b.

Next, suppose instead we had taken $q_1 = -1 - 2i$ so $r_1 = i - 1$. Then $a_2 = 1 + 2i$ and $b_2 = i - 1$. Note $1 + 2i = (-i)(i - 1) + i$, so we can take $q_2 = -i$, $r_2 = i$, so $a_3 = i - 1$ and $b_3 = i$. Clearly $a_3 = 1 \cdot b_3 - 1$, so we can take $a_4 = i$ and $b_4 = -1$. Then Step (2') tells us -1 is a gcd of a and b. (In fact, for the second step we could've also taken $q_2 = 1 - i$ and $r_2 = 1$ which would give 1 as a gcd like in the previous case.)

Finally, suppose we had taken $q_1 = -i$ so $r_1 = -2i$. Then $a_2 = 1 + 2i$ and $b_2 = -2i$. We can write $1 + 2i = (-1)(-2i) + 1$, so we can take $q_2 = -1$, $r_2 = 1$, which gives $a_3 = -2i$ and $b_3 = 1$. Again this gives us a gcd of 1.

This example show that for given a, b the Euclidean algorithm can produce different sequences of a_i's and b_i's, as well result in as different elements of $\mathrm{GCD}(a,b)$, because the division algorithm provides multiple solutions to choose from at each stage. However, in the case that we have a Euclidean algorithm, or even just unique factorization, all gcd's of a and b differ by units, so this algorithm will always output the same final result up to a unit:

Exercise 2.4.4. Prove that if $\mathbb{Z}[\sqrt{-d}]$ has unique factorization, then for $a, b \in \mathbb{Z}[\sqrt{-d}]$ not both zero, $\mathrm{GCD}(a,b)$ is of the form $\{mu : u \text{ is a unit}\}$ for some fixed gcd m of a and b.

Exercise 2.4.5. Use the Euclidean algorithm to compute a gcd of $5 - 5i$ and $3 + 4i$ in $\mathbb{Z}[i]$.

2.5 Unique factorization beyond $\mathbb{Z}$

Now we come to the main results of this chapter beyond the case of $\mathbb{Z}$.

Theorem 2.5.1. *Let $d = 1, 2$. Then $\mathbb{Z}[\sqrt{-d}]$ satisfies the prime divisor property and has unique factorization.*

Proof. Recall from Theorem 2.2.5 that having the prime divisor property implies unique factorization (in fact by Exercise 2.2.3 they are equivalent), so it suffices to prove the prime divisor property. Now that we have a Euclidean algorithm, the proof is similar to that for $\mathbb{Z}$ in Theorem 2.3.3.

Let $a, b \in \mathbb{Z}[\sqrt{-d}]$ be nonzero nonunits. Suppose $p \in \mathbb{Z}[\sqrt{-d}]$ is irreducible and $p \mid ab$ but $p \nmid a$. To prove the prime divisor property, it suffices to show $p \mid b$. Since p is irreducible, its only divisors are of the form u and up where u is a unit of $\mathbb{Z}[\sqrt{-d}]$. Since $p \nmid a$, $up \nmid a$ for any unit u, hence $\mathrm{GCD}(p, a) = \{u : u \text{ is a unit}\}$.

Now, since we have a Euclidean algorithm for $d = 1, 2$, we can apply it to $\{a_1, b_1\} = \{p, a\}$. Then after one pass we have $a_2 = b_1$, $b_2 = r_1 = a_1 - q_1 b_1$ (in the notation of the algorithm in the previous section), i.e., either $\{a_2, b_2\} = \{a, p - q_1 a\}$ or $\{a_2, b_2\} = \{p, a - q_1 p\}$ (basically depending on whether $N(p) \geq N(a)$ or not). Inductively, it follows that at each step in this algorithm, a_i and b_i are $\mathbb{Z}[\sqrt{-d}]$-linear combinations of a and p, i.e., of the form $xa + yp$ for some $x, y \in \mathbb{Z}[\sqrt{-d}]$.

Since we eventually arrive at some a_k being a gcd $u \in \mathrm{GCD}(p, a)$, we have

$$ax + py = u,$$

for some $x, y \in \mathbb{Z}[\sqrt{-d}]$ and some unit u. (Replacing x, y with $u^{-1}x$, $u^{-1}y$, we could assume $u = 1$ if we want.) Multiplying by b gives

$$abx + pby = ub.$$

Since p divides the left hand side, $p \mid ub$, whence $p \mid b$. □

One of the main results of this course will be a proof of Fermat's classification of which numbers n are sums of two squares, i.e., when is $x^2 + y^2 = n$ solvable for $x, y \in \mathbb{Z}$? We will use unique factorization in $\mathbb{Z}[i]$ to prove this, and similarly one can use unique factorization in $\mathbb{Z}[\sqrt{-2}]$ to determine when $x^2 + 2y^2 = n$ has a solution. We will get to this in Chapter 4 after first treating modular arithmetic in Chapter 3.

The reason for this ordering is that I'm trying to roughly organize the chapters by topics, and wanted to prove the fundamental theorem of arithmetic before doing modular arithmetic, as we will talk about factorization problems related to cryptography in that chapter. But we will also use modular arithmetic in the determination of numbers which are sums of 2 squares, so that needed to go before we get our main application of Theorem 2.5.1. However, since we worked so long to get this theorem, I'd like to show you know how to get a nice application right now. Here is one, though I leave some details to the exercises.

Definition 2.5.2. *Let $d \in \mathbb{Z}$ and $a, b \in \mathbb{Z}[\sqrt{d}]$. We say a and b are* **relatively prime** *(or* **coprime***) if 1 is a gcd of a and b.*

Proposition 2.5.3. *The only solution to $y^3 = x^2 + 2$ in $\mathbb{N}$ is $(x, y) = (5, 3)$.*

Proof. Suppose

$$y^3 = x^2 + 2 = (x + \sqrt{-2})(x - \sqrt{-2}).$$

It is not hard to show that $x + \sqrt{-2}$ and $x - \sqrt{-2}$ are relatively prime in $\mathbb{Z}[\sqrt{-2}]$ (first exercise below). Since their product is a cube (i.e., of the form a^3 for some $a \in \mathbb{Z}[\sqrt{-2}]$), then both $x + \sqrt{-2}$ and $x - \sqrt{-2}$ are cubes by unique factorization in $\mathbb{Z}[\sqrt{-2}]$ (second exercise below). Write

$$x + \sqrt{-2} = (m + n\sqrt{-2})^3 = m^2 - 6mn^2 + (3m^2n - 2n^3)\sqrt{-2}$$

for some $m, n \in \mathbb{Z}$. Hence

$$x = m^3 - 6mn^2\, 1 = n(3m^2 - 2n^2).$$

From the second equation, we have $n = \pm 1$ and $3m^2 - 2n^2 = 3m^2 - 2 = 1$, so $m = \pm 1$ and $x = \pm 5$. □

Exercise 2.5.1. Suppose $x, y \in \mathbb{N}$ such that $y^3 = x^2 + 2$. This is how to prove $x + \sqrt{-2}$ and $x - \sqrt{-2}$ are coprime in $\mathbb{Z}[\sqrt{-2}]$.
(i) Show x must be odd.
(ii) Show for any odd $x \in \mathbb{Z}$, $x+\sqrt{-2}$ and $x-\sqrt{-2}$ are coprime in $\mathbb{Z}[\sqrt{-2}]$. (*Suggestion:* Try adding and subtracting these two quantities.)

Exercise 2.5.2. Let $d = 1, 2$. Use Theorem 2.5.1 to show that if ab is a cube in $\mathbb{Z}[\sqrt{-d}]$ and a and b are relatively prime, then a and b are cubes in $\mathbb{Z}[\sqrt{-d}]$.

Now that we've seen some concrete utility of unique factorization in quadratic rings, you might wonder for what other quadratic rings one gets unique factorization. You might be worried from our proof of the division algorithm that we don't actually get unique factorization for any other imaginary quadratic rings. However, that's not exactly true.

Let's recall how various properties of quadratic rings are related:

$$\begin{aligned} &\text{division property} \implies \\ &\text{Euclidean algorithm} \implies \\ &\text{prime divisor property} \iff \\ &\text{unique factorization} \end{aligned}$$

So we while we suggested that $\mathbb{Z}[\sqrt{-d}]$ does not have the division property for $d \geq 3$, that doesn't mean we need that or the Euclidean algorithm to have prime factorization, but only that it's a useful tool for proving unique factorization when $d = 1, 2$.

Let's think about the next example, $d = 3$. We saw in Exercise 2.4.2 that $\mathbb{Z}[\sqrt{-3}]$ does not have the division property, and we also saw in Exercise 2.2.2 that $\mathbb{Z}[\sqrt{-3}]$ does not have unique factorization, namely

$$4 = 2 \cdot 2 = (1 + \sqrt{-3})(1 - \sqrt{-3})$$

are two distinct factorizations of 4 into irreducibles of $\mathbb{Z}[\sqrt{-3}]$. However, in some sense, the problem with this example is simply that $\mathbb{Z}[\sqrt{-3}]$ does not have enough elements, and

this can be resolved by passing to the Eisenstein integers $\mathbb{Z}[\zeta_3] = \mathbb{Z}[\frac{1+\sqrt{-3}}{2}]$, which contains $\mathbb{Z}[\sqrt{-3}]$. Namely, in $\mathbb{Z}[\zeta_3]$ we can factor

$$2 = \frac{1+\sqrt{-3}}{2} \cdot (1 - \sqrt{-3}),$$

so the above factorizations of 4 in $\mathbb{Z}[\sqrt{-3}]$ resolve into the factorizations

$$4 = (\frac{1+\sqrt{-3}}{2} \cdot (1-\sqrt{-3}))^2 = \left(\frac{1+\sqrt{-3}}{2} \cdot (1-\sqrt{-3})\frac{1+\sqrt{-3}}{2}\right)(1-\sqrt{-3}).$$

Indeed one has unique factorization in $\mathbb{Z}[\zeta_3]$, and one can prove it in a similar manner to above. The key step is the division property:

Exercise 2.5.3. Show $\mathbb{Z}[\zeta_3]$ has the division property. (*Suggestion:* First think about the rectangles we constructed for $\mathbb{Z}[\sqrt{-3}]$ in the proof of Theorem 2.4.5. Then think about what multiples you get inside these rectangles if you can multiply by $\frac{1+\sqrt{-3}}{2}$.)

Theorem 2.5.4. *The Eisenstein integers $\mathbb{Z}[\zeta_3]$ have unique factorization.*

Proof. Once one has the division property, one gets a Euclidean algorithm, and the proof is the same as that for $\mathbb{Z}[i]$ and $\mathbb{Z}[\sqrt{-2}]$. □

Unique factorization for $\mathbb{Z}[\zeta_3]$ is useful for proving Fermat's Last Theorem for $n = 3$, i.e., $x^3 + y^3 = z^3$ has no nontrivial solutions (i.e., no solutions with x, y, z all nonzero). We'll discuss this in Chapter 6.

Exercise 2.5.4. Even though $\mathbb{Z}[\zeta_3]$ has unique factorization, show that there exist $a, b \in \mathbb{Z}[\zeta_3]$ which are relatively prime with ab a cube but neither a nor b are cubes. What is the difference between this situation and Exercise 2.5.2?

Now that we've seen $\mathbb{Z}[\zeta_3]$ can resolve failure of unique factorization in $\mathbb{Z}[\sqrt{-3}]$, you might wonder if you can do this for other rings, e.g., $\mathbb{Z}[\sqrt{-5}]$ or $\mathbb{Z}[\sqrt{-7}]$? It turns out you can in one but not the other. For $\mathbb{Z}[\sqrt{-3}]$ this worked because we could adjoin "quadratic integer" $\frac{1+\sqrt{-3}}{2}$, and then the geometry works out nice in Exercise 2.5.3.

Regarding the quadratic integer terminology, what it means in general for a complex number z to be an **algebraic integer** is that it is a root of some monic polynomial $z^n + c_{n-1}z^{n-1} + \cdots c_1 z + c_0$ for some $c_i \in \mathbb{Z}$. E.g., since ζ_n^j satisfies $z^n - 1 = 0$, each root of unity ζ_n^j is an algebraic integer. So are sums, differences and products of algebraic integers (the set of all algebraic integers forms a ring). In particular $\zeta_3 = \frac{-1+\sqrt{-3}}{2}$ is an algebraic integer, and so is $1 + \zeta_3 = \frac{1+\sqrt{-3}}{2}$.

We say $d \in \mathbb{Z}$ is **squarefree** if $m \in \mathbb{N}$ with $m^2 | d$ implies $m = 1$, i.e., d is not nontrivially divisible by any squares. Note that if $d \in \mathbb{Z}$ with $d = d_0 m^2$ where d_0 is squarefree and $m \in \mathbb{N}$, then $\sqrt{d} = m\sqrt{d_0}$ so

$$\mathbb{Z}[\sqrt{d}] = \left\{a + bm\sqrt{d_0} : a, b \in \mathbb{Z}\right\} = \mathbb{Z}[m\sqrt{d_0}]$$
$$\mathbb{Q}(\sqrt{d}) = \left\{a + bm\sqrt{d_0} : a, b \in \mathbb{Q}\right\} = \mathbb{Q}(\sqrt{d_0}).$$

So when we want to work with quadratic rings which are as big as possible (and thus increasing the likelihood of unique factorization, admittedly for reasons I haven't completely explained), we may as well restrict to d squarefree. Moreover, one gets all quadratic fields $\mathbb{Q}(\sqrt{d})$ by restricting to squarefree d.

Definition 2.5.5. *Let $d \in \mathbb{Z}$ be squarefree, $d \neq 0, 1$. The* **ring of integers** *of $\mathbb{Q}(\sqrt{d})$ is*

$$\mathcal{O}_d = \begin{cases} \mathbb{Z}[\frac{1+\sqrt{d}}{2}] = \left\{a + b \cdot \frac{1+\sqrt{d}}{2} : a, b \in \mathbb{Z}\right\} & \text{if } d \equiv 1 \bmod 4, \\ \mathbb{Z}[\sqrt{d}] & \text{else.} \end{cases}$$

Exercise 2.5.5. Show each $\mathcal{O}_d$ is a ring.

There is a general result that says that (for d any non-square) $\mathbb{Z}[\sqrt{d}]$ can only have a Euclidean algorithm or unique factorization if $\mathbb{Z}[\sqrt{d}]$ is a full ring of integers $\mathcal{O}_d$. Necessarily, d is both squarefree and $d \not\equiv 1 \bmod 4$.

We now give a brief summary of what is known. Gauss, in his *Disquistiones Arithmeticae* from 1801 (when he was only 23), which was a major milestone in number theory, found that imaginary quadratic rings of integers $\mathcal{O}_d$ ($d < 0$) have unique factorization when

$$d = -1, -2, -3, -7, -11, -19, -43, -67, -163$$

and conjectured there are no others. On the other hand, Gauss also conjectured that there are infinitely many real quadratic rings of integers $\mathcal{O}_d$ ($d > 0$) with unique factorizations.

Gauss's conjecture on imaginary quadratic rings was proved in the 1960's by Heegner (a high school teacher) and Stark, and independently by Baker. Gauss's conjecture on real quadratic rings is still one of the major open problems in number theory, despite that unique factorization seems extremely common for real quadratic $\mathcal{O}_d$. (More precisely, Cohen and Lenstra conjectured $\mathcal{O}_d$ should have unique factorization for more than 75% of squarefree $d > 1$.)

In any case, we see that having unique factorization is a rather special property for rings we care about in number theory—e.g., it only happens for finitely many imaginary quadratic rings. We also remark that it only happens finitely many cyclotomic rings $\mathbb{Z}[\zeta_n]$ (the largest of which is with $n = 90$). Much of algebraic number theory was developed as an attempt to understand unique factorization and how to work in number rings when unique factorization fails. There are three main ways to "resolve" nonunique factorization: (i) Gauss's theory of binary quadratic forms for quadratic rings; (ii) Kummer's theory of ideal numbers (basically, work in a larger ring S than your given ring R, where unique factorization still might not hold, but that any factorization of R into irreducibles of S is essentially unique); (iii) Dedekind's ideal theory (a generalization of Kummer's ideal numbers, where one work with certain sets instead of number). Of these (iii) is by far the most common theory to use regarding factorization in number rings in modern number theory. Now we're not studying any of these in this class—(ii) and (iii) are more advanced than what we will do in this course, whereas (i) goes in a different direction.

Going back to our musings on which imaginary quadratic rings have a Euclidean algorithm or unique factorization, the proof of Gauss's conjecture in the imaginary case tells

us that $\mathbb{Z}[\sqrt{-d}]$ ($d > 0$) only has unique factorization when $d = 1, 2$, the cases we proved. To see this, note the other cases where $\mathcal{O}_{-d}$ has unique factorization with $d > 0$ fall in the situtation $-d \equiv 1 \bmod 4$, i.e., $d \equiv 3 \bmod 4$, so $\mathbb{Z}[\sqrt{-d}] \neq \mathcal{O}_{-d}$, and use the "general result" mentioned above. Consequently, the only cases where $\mathbb{Z}[\sqrt{-d}]$ can have a Euclidean algorithm is when $d = 1, 2$, and we proved in these cases it does.

One does not see this by just looking at the rings $\mathbb{Z}[\sqrt{-d}]$, but having a Euclidean algorithm really is more special than just having unique factorization. For the imaginary quadratic rings of integers $\mathcal{O}_{-d}$, they only have Euclidean algorithms if $d = -1, -2, -3, -7$, or -11. Thus the last 4 cases on Gauss's imaginary quadratic list have unique factorization but no Euclidean algorithm. Among real quadratic rings $\mathcal{O}_d$, we only have a Euclidean algorithm (with respect to the norm[4]) when

$$d = 2, 3, 5, 6, 7, 11, 13, 17, 19, 21, 29, 33, 37, 41, 57, 73.$$

[4]There is a more general notion of a ring having a "Euclidean algorithm," namely one can measure size not just using the norm, but possibly using another function. It is conjectured that infinitely many real quadratic rings $\mathcal{O}_d$ have a Euclidean algorithm in this more general sense, but still there are $\mathcal{O}_d$ which have unique factorization but do not have a generalized Euclidean algorithm.

Chapter 3

Modular Arithmetic

In this chapter, we'll look at some applications of modular arithmetic, i.e., applications of the rings $\mathbb{Z}/n\mathbb{Z}$ to number theory. In particular, we'll get applications to divisibility tests, necessary conditions for solutions of various Diophantine equations (including non-solvability results), as well as an application to modern cryptography. For some of these applications, we will need a deeper understanding of the arithmetic structure of $\mathbb{Z}/n\mathbb{Z}$, such as knowing which elements of $\mathbb{Z}/n\mathbb{Z}$ have a multiplicative inverse and when $\mathbb{Z}/n\mathbb{Z}$ is a field. For this, will take another little detour into abstract algebra with the notion of groups. (Thus we will have hit the 3 main types of algebraic structures covered in an abstract algebra course: groups, rings, and fields—albeit mainly restricted to the commutative setting.)

3.1 Divisibility criteria

One of the most basic applications of modular arithmetic is to obtaining the classic divisibility tests based on the decimal (base 10) representation of n.

Proposition 3.1.1. *Let $n \in \mathbb{N}$. Then n is divisible by 2, 5 or 10 if and only if its last digit is. Similarly, n is divisible by 4, 25 or 100 if and only if the integer consisting of its last two digits is.*

If $n < 10$, we interpret the last two digits to just mean n (i.e., write n in decimal with a preceding 0).

Proof. Write $a_d a_{d-1} \cdots a_1 a_0$ as the base 10 representation of n, i.e., $0 \leq a_i \leq 9$ and

$$n = 10^d a_d + 10^{d-1} a_{d-1} + \cdots + 10^1 a_1 + 10^0 a_0.$$

If $m = 2$, 5 or 10 then $m \mid 10$ so $n \equiv a_0 \bmod m$. Hence $m|n$ (i.e., $n \equiv 0 \bmod m$) if and only if $m \mid a_0$.

If $m = 4$, 25 or 100, then, then $10^j \equiv 0 \bmod m$ for $j \geq 2$, so $n \equiv 10a_1 + a_0 \bmod m$. So again, $m \mid n$ if and only if $m \mid (10a_1 + a_0)$. □

The above argument can be written easily enough without modular arithmetic, but the the standard divisibility tests for 3 and 9 are really much more transparent with modular arithmetic.

Proposition 3.1.2. *Let $n \in \mathbb{N}$. Then n is divisible by 3 or 9 if and only if the sum of its digits is.*

Proof. Let $a_d a_{d-1} \cdots a_1 a_0$ be the base 10 representation of n, i.e., $0 \leq a_i \leq 9$ and

$$n = 10^d a_d + 10^{d-1} a_{d-1} + \cdots + 10^1 a_1 + 10^0 a_0.$$

Let $m = 3$ or 9. Since $10 \equiv 1 \bmod m$, we have $10^j \equiv 1^i \equiv 1 \bmod m$ for any i. Hence

$$n \equiv a_d + a_{d-1} + \cdots + a_1 + a_0 \bmod m.$$

Again, this means $m \mid n$ if and only if $m \mid \sum a_i$. □

Exercise 3.1.1. Let $n \in \mathbb{N}$. Show n is divisible by 11 if and only if the alternating sum of its digits is. (E.g., by the alternating sum of the digits of the number 12345, we mean $1 - 2 + 3 - 4 + 5$.)

We can use the same idea to give divisibility criteria in terms of representations of numbers in other bases. Here is a simple example which is similar to the last problem.

Exercise 3.1.2. Consider the binary expansion of $n \in \mathbb{N}$, which consists of a string of *bits* ("binary digits"). Show that n is divisible by 3 if and only if the alternating sum of its bits is.

From above we have tests for divisibility of n in terms of its digits for dividing by any number up to 10, except for 7 and 8. We didn't state one explicitly for divisibility by 6, but clearly you can just test for divisibility by 2 and by 3 thanks to unique factorization, or more directly the prime divisor property. (Think about why the prime divisor property is relevant.) You can also use a simple test for 8, generalizing the ones for 2 and 4, which is a special case of the following:

Exercise 3.1.3. Let $k, n \in \mathbb{N}$. Show n is divisible by 2^k if and only if the number consisting of just the last k digits of n is. Moreover, show that looking at the last $k - 1$ digits does not suffice to determine divisibility by 2^k.

Probably you knew about most of these divisibility tests already (though maybe you didn't know how to prove some of them). On the other hand, you probably don't know a divisibility test for 7, and that's because such a test is more complicated, though you can still write one down:

Exercise 3.1.4. Let $n \in \mathbb{N}$. Devise a test to determine if n is divisible by 7 or not, based on looking at certain combinations of digits.

3.2 Applications to Diophantine equations

Recall from the introduction that one of standard descriptions of number theory is the study of Diophantine equations. To be formal, here is a proper definition:

Definition 3.2.1. *A* **Diophantine equation** *is an equation of the form* $f(x_1, \dots, x_n) = g(x_1, \dots, x_n)$*, where* $x_1, \dots, x_n$ *are variables in* $\mathbb{Z}$ *and* f, g *are polynomials with coefficients in* $\mathbb{Z}$*.*

Note such an equation is equivalent to the equation $F(x_1, \dots, x_n) = 0$ where F is the polynomial $f - g$, so when we discuss Diophantine equations it suffices to assume the equation is in the form $F(x_1, \dots, x_n) = 0$.

Since we take $x_1, \dots, x_n$ to be variables in $\mathbb{Z}$, by a solution to a Diophantine equation $F(x_1, \dots, x_n) = 0$ we mean a solution with each $x_i \in \mathbb{Z}$, which we call a **solution over** $\mathbb{Z}$.[1] Thus solving Diophantine equations is equivalent to finding integer roots of polynomials with integer coefficients.

To remind you where we're going, the following families of Diophantine equations—all of which were discussed in the introduction—are the main Diophantine equations we are focused on in this course.

(1) $x^2 + y^2 = n$ (which numbers are sums of 2 squares? and to a lesser extent, which numbers are of the form $x^2 + dy^2$? see Chapter 4)

(2) $x^2 + y^2 + z^2 + w^2 = n$ (which numbers are sums of 4 squares? see Section 4.6)

(3) $x^2 - dy^2 = 1$ (Pell's equation, related to finding rational approximations for $\sqrt{d}$; see Chapter 5)

(4) $x^3 + y^3 = z^3$ (we'll also say a bit more generally about $x^n + y^n = z^n$, the subject of Fermat's Last Theorem; see Chapter 6[2])

Here we regard n and d as constants in these equations. The goal is to determine when these equations have solutions and, if possible, describe all solutions or explain how to find all solutions. We already treated the simple case of linear Diophantine equations $ax + by = c$ in 2 variables in Proposition 2.3.1, where a, b, c are constants.

For instance, for the first family of equations above, $x^2 + y^2 = n$, we mainly want to do two things: (i) for n such that a solution exists prove one exists, and (ii) for n such that no solution exists prove there is no solution. In this case, for given n, it is not hard to determine solutions algorithmically—one can simply check values of $x^2 + y^2$ for $0 \le x, y \le \sqrt{n}$ similar to the proof of part (1) of Proposition 1.5.5. There are of course more efficient algorithms, but we will not focus on algorithmic aspects too much in this course. While there's no simple formula in general (in terms of n) for solutions to $x^2 + y^2 = n$, another thing one can do is

[1] Technically, the phrasing "a solution *in* $\mathbb{Z}$" would mean that the solution to the equation is a single integer in $\mathbb{Z}$, rather than a tuple of integers, so I will try to say a solution *over* $\mathbb{Z}$ when there is more than one variable, but forgive me if I make a *faux pas*. On the other hand, I may say "integer solution" or "integral solution" for a solution over $\mathbb{Z}$ which is not a single integer in $\mathbb{Z}$ but a tuple in $\mathbb{Z}^n$. (This can be grammatically justified by calling $\mathbb{Z}^n$ the set of integer or integral points in $\mathbb{R}^n$ or $\mathbb{C}^n$.)

[2] Footnote from the future: Yeah, this didn't happen.

count the number of solutions, which is a refinement of just determining whether solutions exist or not (i.e., determine when the count is positive versus zero). We won't focus too much on actually counting the number of solutions in this course, but we'll say a little about this also. (At a crude level, we've already noted that the number of solutions to (1), and similarly (2), must be finite, whereas the number of solutions to (3) can be infinite.)

The easiest way to show that a Diophantine equation has a solution is exhibit a solution. Recall, for $ax + by = c$, we didn't give a formula for solutions x, y but rather an algorithm for finding solutions x, y when they exist, which the most practical thing one can hope for as there are typically no simple formulas for solutions to Diophantine equations. For the above equations, one needs to work harder to show solutions exist.

On the other hand, much of the time there is an easy way to show solutions don't exist. That comes via modular arithmetic.

Proposition 3.2.2. *Let $F(x_1, \dots, x_n) = 0$ be a Diophantine equation. If this equation has a solution, then*

$$F(x_1, \dots, x_n) \equiv 0 \bmod m, \tag{3.2.1}$$

has a solution for all $m \in \mathbb{N}$.

The equation (3.2.1) is called the **reduction mod** m of $F(x_1, \dots, x_n) = 0$, and we may view it as an equation in n variables in $\mathbb{Z}/m\mathbb{Z}$.

Proof. Suppose $x_1, \dots, x_n \in \mathbb{Z}$ such that $F(x_1, \dots, x_n) = 0$. Then $m \mid F(x_1, \dots, x_n)$ for all $m \in \mathbb{N}$ (in fact for all $m \in \mathbb{Z}$ if one wants). □

The point is that it is often easy to show an equation mod m doesn't have any solutions. Algorithmically, certainly it's very simple: there are only m possibilities for $x_1, \dots, x_n$ regarded as elements of $\mathbb{Z}/m\mathbb{Z}$, so at most we need to compute $F(x_1, \dots, x_n)$ mod m for a total of m^n possible inputs.

Remark 3.2.3. It is *not* true that the converse of the proposition holds. Namely, there are Diophantine equations which have solutions mod m for all m, but do not have solutions over $\mathbb{Z}$. A couple of famous examples are $x^2 + y^2 + z^2 + w^2 = -1$ and $3x^3 + 4y^3 + 5z^3 = 0$. The problem in some sense is that while these have solutions mod m for all m, you can't choose the solutions in a compatible way to "lift" them to solutions over $\mathbb{Z}$. One of the major themes in modern number theory is to study to what extent you can lift solutions mod m to solutions over $\mathbb{Z}$. To read more about this, look up the *local-global principle*. One particularly fascinating situation is the family of equations of the form $x^2 + dy^2 = n$ (here $d > 0$). It turns out that the problem of lifting solutions mod m to solutions over $\mathbb{Z}$ is related to the failure of unique factorization in $\mathbb{Z}[\sqrt{-d}]$. In particular, if one has unique factorization in $\mathbb{Z}[\sqrt{-d}]$ (or if unique factorization doesn't fail "too badly") then $x^2 + dy^2 = n$ has a solution over $\mathbb{Z}$ if and only if it does mod m for all m and $n \geq 0$. On the other hand, this is not true for $d = 23$, where unique factorization fails "sufficiently badly." In particular, $x^2 + 23y^2 = 41$ has a solution mod m for all m but no integer solution.

Example 3.2.1. Let $n \in \mathbb{Z}$ and $f(x) = x^2 + x$. If n is odd, then $f(x) = n$ has no solution. To see this, look at the equation mod 2, which is simply $x^2 + x \equiv n \bmod 2$ Now either

$x \equiv 0 \bmod 2$ or $x \equiv 1 \bmod 2$. In either case, we see $x^2 + x \equiv x(x+1) \equiv 0 \bmod 2$, whence n must be even to get a solution.

Of course we could just make this argument in terms of even and odd numbers, but the benefit of this language of modular arithmetic is that it greatly generalizes what you can easily do just by thinking in terms of evens and odds. For instance, consider $x^2 + x \bmod 3$. This is 0 when $x \equiv 0, 2 \bmod 3$ and 2 when $x \equiv 1 \bmod 3$, so $x^2 + x \equiv 1 \bmod 3$ has no solutions. Thus we can conclude that any integer n of the form $x^2 + x$ ($x \in \mathbb{Z}$) must be even and not $\equiv 1 \bmod 3$, i.e., $6 \mid n$ or $6 \mid (n-2)$.

Exercise 3.2.1. Determine the possibilities for $x^2 + x \bmod 5$ and $x^2 + x \bmod 7$. Using this, and the previous example, completely determine which $0 \leq n \leq 20$ are of the form $x^2 + x$.

Since our next example is important in determining which numbers are sums of two squares, one of the main goals of the course, we elevate its status to a proposition.

Proposition 3.2.4. *Let $n \in \mathbb{N}$. If $n \equiv 3 \bmod 4$, then n is not a sum of 2 (integer) squares, i.e., $x^2 + y^2 = n$ has no solution over $\mathbb{Z}$.*

Note this criterion provides a great speed-up to the algorithmic approach to looking for solutions to $x^2 + y^2 = n$. We can just first check $n \bmod 4$ (for which it suffices to check the last 2 digits), and if we get 3 mod 4 stop. Of course if n is not 3 mod 4, we still need to look for solutions.

The proof requires the notion of squares mod n. We also say an integer $a \in \mathbb{Z}$ is a **square mod** n if $a + n\mathbb{Z}$ is a square in $\mathbb{Z}/n\mathbb{Z}$, i.e., $a \equiv x^2 \bmod n$ for some $x \in \mathbb{Z}$. Otherwise, we say a is a **non-square mod** n. Since being a square (or non-square) mod n only depends upon the equivalence class mod n, we will sometimes think of the squares (or non-squares) mod n as elements of $\mathbb{Z}/n\mathbb{Z}$.

Example 3.2.2. Let $n \geq 2$. Then $0^2 \equiv 0 \bmod n$ and $1^2 \equiv 1 \bmod n$, so there are always at least 2 squares mod n (thought of as elements of $\mathbb{Z}/n\mathbb{Z}$). On the other hand there are at most n, as there are n elements of $\mathbb{Z}/n\mathbb{Z}$. In particular, all numbers are squares mod 2.

Example 3.2.3. Note that $0^2 \equiv 2^2 \equiv 0 \bmod 4$ and $1^2 \equiv 3^2 \equiv 1 \bmod 4$. Put another way, the square of an even number is 0 mod 4 and the square of an odd number is 1 mod 4. Hence the squares mod 4 are simply 0 and 1 (thought of as elements of $\mathbb{Z}/4\mathbb{Z}$), and 2 and 3 (as elements of $\mathbb{Z}/4\mathbb{Z}$, i.e., technically $2 + 4\mathbb{Z}$ and $3 + 4\mathbb{Z}$) are non-squares mod 4.

Example 3.2.4. Note $2^2 \equiv (-1)^2 \equiv 1 \bmod 3$, so the elements 0 and 1 of $\mathbb{Z}/3\mathbb{Z}$ are squares and $-1 \equiv 2 \bmod 3$ is a non-square.

Proof of proposition. Since the squares mod 4 are 0 and 1, we have one of the following possibilities for $x, y \in \mathbb{Z}$:

$$x^2 + y^2 \equiv \begin{cases} 0 + 0 \equiv 0 \bmod 4 \\ 1 + 0 \equiv 1 \bmod 4 \\ 0 + 1 \equiv 1 \bmod 4 \\ 1 + 1 \equiv 2 \bmod 4. \end{cases}$$

Thus the sum of 2 squares is never 3 mod 4. □

We remark one can also formulate the proposition as a divisibility statement: the: for any x, y, $x^2 + y^2 - i$ is divisible by 4 for some $i = 0, 1, 2$, i.e., $f(x, y) = (x^2 + y^2)(x^2 + y^2 - 1)(x^2 + y^2 - 2)$ is always divisible by 4. Here are some similar, rather well known, examples.

Exercise 3.2.2. Show $x^2 + 2y^2 = n$ has no solution over $\mathbb{Z}$ if $n \equiv 5, 7 \bmod 8$.

Exercise 3.2.3. Show $x^2 + 3y^2 = n$ has no solution over $\mathbb{Z}$ if $n \equiv 2 \bmod 3$.

Exercise 3.2.4. Show that $n \in \mathbb{N}$ is not a sum of 3 (integer) squares if $n \equiv 7 \bmod 8$.

Exercise 3.2.5. Show that $n \in \mathbb{N}$ is not a sum of two (integer) cubes if $n \equiv 4, 5 \bmod 9$.

More generally than just getting non-existence of solutions to certain Diophantine equations, we can also obtain necessary conditions for solutions to Diophantine equations. This is useful for (i) helping to find solutions when they exist, and (ii) as an intermediary step for proving the non-existence of solutions when they don't exist. Here's a simple example of this technique.

Proposition 3.2.5. *Suppose $x, y, z, w \in \mathbb{Z}$ such that $x^2 + y^2 + z^2 = w^2$. If w is odd, then exactly one of x, y, z is odd. If w is even, then all of x, y, z are even.*

Proof. Recall that the squares mod 4 are 0 and 1. Note that if w is odd, then an odd number of x, y, and z are odd, i.e., an odd number of x^2, y^2 and z^2 are 1 mod 4. If all three are, then $x^2 + y^2 + z^2 \equiv 3 \bmod 4$, but $w^2 \equiv 1 \bmod 4$. Hence exactly one of x, y and z is odd.

The argument for w even is similar, and we leave it to the reader. □

Note even though the original statement is only about the parity of solutions, looking at things mod 2 is not sufficient to prove this statement, as all numbers are squares mod 2. For instance, when w is odd, then looking at parities only tells you that an odd number of x, y and z must be odd.

Example 3.2.5. Now let's use the above proposition to determine all solutions to $x^2 + y^2 + z^2 = 9$ with $x, y, z \in \mathbb{N}$. We know exactly one of x, y and z must be odd. So two of them must be at least 2, which forces the other to be 1. Consequently all solutions over $\mathbb{N}$ are $(2, 2, 1)$, $(2, 1, 2)$ and $(1, 2, 2)$.

Exercise 3.2.6. Determine if $x^2 + y^2 + z^2 = 25$ has any solutions with $x, y, z \in \mathbb{N}$. If so, find all solutions.

Exercise 3.2.7. Determine if $x^2 + y^2 + z^2 = 64$ has any solutions with $x, y, z \in \mathbb{N}$. If so, find all solutions.

In Chapter 6, we'll see how we can use this technique to make a little progress on Fermat's last theorem, however the only known ways to prove Fermat's last theorem use much more advanced machinery than just simple considerations mod m.

3.3 Groups and invertibility mod n

To go a bit further with applications of modular arithmetic, we need to understand some things about the multiplicative structure of $\mathbb{Z}/n\mathbb{Z}$. In this section, except where noted otherwise, we assume $n > 1$.

Definition 3.3.1. *We say $a \in \mathbb{Z}$ is* **invertible** *mod n if $a + n\mathbb{Z}$ has a multiplicative inverse in $\mathbb{Z}/n\mathbb{Z}$, i.e., if there exists $b \in \mathbb{Z}$ such that $ab \equiv 1 \bmod n$. In this case b is called a* **(multiplicative) inverse** *of a mod n.*

Note that this only depends on the congruence class, i.e., if $a \equiv a' \bmod n$, then a is invertible mod n if and only if a' is, and the inverse only depends on the congruence class as well. As with the notion of squares mod n, we sometimes think of inverses mod n as integers, and sometimes as elements of $\mathbb{Z}/n\mathbb{Z}$, depending on which is more convenient.

The notion of invertibility can also be phrased in terms of Diophantine equations mod n: a is invertible mod n if and only if $ax \equiv 1 \bmod n$ has a solution in $\mathbb{Z}/n\mathbb{Z}$.

The invertible elements of $\mathbb{Z}/n\mathbb{Z}$ (or more generally a ring) will give us an algebraic structure known as a group.

Definition 3.3.2. *Let G be a set with a binary operation $\cdot$. We say $(G, \cdot)$ (or just G if the operation is understood) is a* **group**, *if the following three properties hold:*

(1) $\cdot$ is associative: $(g \cdot h) \cdot k = g \cdot (h \cdot k)$ for all $g, h, k \in G$;

(2) there is an identity $1 \in G$ such that $1 \cdot g = g \cdot 1 = g$ for all $g \in G$;

(3) every $g \in G$ has an **inverse** *g^{-1} such that $g^{-1} \cdot g = g \cdot g^{-1} = 1$;*

If G is a group which also satisfies

(4) $\cdot$ is commutative: $g \cdot h = h \cdot g$ for all $g, h \in G$,

then we say $(G, \cdot)$ (or just G) is an **abelian group**.

When the operation is understood, we typically write gh for $g \cdot h$, and this notation is called **multiplicative notation**. However, for some abelian groups, the operation $\cdot$ will be written as $+$, which is called **additive notation**. In the case of additive notation, we denote the identity by 0 instead of 1, and the inverse of g by $-g$ instead of g^{-1}. Accordingly, an **additive group** will mean an abelian group in additive notation, and a **multiplicative group**.

The reason for these conventions should be clear from following simple examples (the proofs are easy, and you may fill them in for yourself).

Example 3.3.1. $(\mathbb{Z}, +)$ is an additive (abelian) group. So is $(\mathbb{Q}, +)$ and $(\mathbb{R}, +)$, or more generally $(R, +)$ where R is any ring. In all cases, the identity of the group is the zero element 0 of the ring, and the inverse of any a in the ring is $-a$. (Our notation for 0, $+$ and $-$ in a ring R is consistent with the additive notation for the group $(R, +)$.) On the other hand, $(\mathbb{N}, +)$ is not a group as it does not have the identity or (additive) inverses.

Example 3.3.2. $(\mathbb{Q}^\times, \times) = \mathbb{Q} - \{0\}$ is an infinite abelian multiplicative group. So is $\mathbb{R}^\times$ and $\mathbb{C}^\times$. We will generalize these examples (with proof) to an arbitrary ring below. Similarly, the positive rational (or reals) also are. In all cases, the identity is the integer 1 and the inverse of any element x is $x^{-1} = \frac{1}{x}$.

Lemma 3.3.3. *Let G be a group. Then there is a unique identity, and each $g \in G$ has a unique inverse.*

Proof. You already proved that any binary operation has at most 1 identity (Exercise 1.2.4), so the identity of G is unique. Now let $g \in G$ and suppose $h, h' \in G$ such that h and h' are inverses of g. Then on one hand $hgh' = (hg)h' = 1 \cdot h' = h'$, but also $hgh' = h(gh') = h \cdot 1 = h$, whence $h = h'$. □

We say a group $(G, \cdot)$ is **finite** if the set G is finite. The finite abelian groups are in some sense the simplest class of groups and have a simple characterization. If G is a finite group with n elements, we say it has **order** n, and write $|G| = n$.

Here are some more examples, mostly without proofs.

Example 3.3.3. $(\mathbb{Z}/n\mathbb{Z}, +)$ is an finite abelian group of order n.

Example 3.3.4. (n-th roots of unity) Recall the n-th roots of unity $\mu_n = \{e^{2\pi ik/n} : 0 \leq k < n\}$. Then, with the standard multiplication, μ_n is a finite abelian group of order n (see exercise below).

We remark that the group μ_n has the same structure (is "isomorphic" to) $(\mathbb{Z}/n\mathbb{Z}, +)$, the only difference being one group is written with multiplicative notation and one with additive notation. (Recall the pictures of $\mathbb{Z}/n\mathbb{Z}$ and μ_n as n points around a circle.) Precisely, if we write down the operation table for $(\mathbb{Z}/n\mathbb{Z}, +)$, with elements represented as $0, 1, \dots, n-1$ in the obvious way, and change each element label i to ζ_n^i and relabel our operation $+$ for $\mathbb{Z}/n\mathbb{Z}$ to $\cdot$, we get exactly the multiplication table for μ_n.[3]

Exercise 3.3.1. Prove μ_n is a group under multiplication. Write down the multiplication table when $n = 3$ and check it looks the same as the addition table for $\mathbb{Z}/3\mathbb{Z}$.

Example 3.3.5. (dihedral groups) Fix $n > 2$. Let P be a regular polygon with n vertices. The the set of automorphisms of P, namely the rotations and reflections which map P to itself, form a finite *non-abelian* group of order $2n$ called the **dihedral group** D_{2n}, where the operation is composition.

Example 3.3.6. (symmetric groups) Let S_n be the set of permutations of $\{1, 2, \dots, n\}$. Then S_n is a finite group of order $n!$ with the composition operation, called the **symmetric group** on n elements. It is non-abelian for $n > 2$.

Note that in the groups in the previous two examples can be naturally thought of as the symmetries of some object—D_{2n} is the set of geometric symmetries of a regular n-gon in a plane, and S_n is the set of "combinatorial" symmetries of a set of size n (though one can also realize S_n geometrically, e.g., as the symmetries of the standard basis of $\mathbb{R}^n$). The standard way of thinking about what the notion of a group represents is the notion of the symmetries of some object: given two symmetries g and h one can compose them to get a new symmetry $g \cdot h$; this composition is associative, the "do nothing" symmetry is the identity, and each symmetry can be applied in reverse giving an inverse. (Historically, group theory was developed to study permutations of roots of polynomials by Galois and others. The term "abelian" is in honor of Niels Abel, who proved that the "Galois group" of a polynomial being commutative means the roots of that polynomial can be found with radicals).

Example 3.3.7. Let $\mathrm{GL}_n(\mathbb{R})$ denote the set of $n \times n$ invertible matrices with real entries. From linear algebra, being invertible simply means the determinant is nonzero. Then $\mathrm{GL}_n(\mathbb{R})$ forms a group with respect to matrix multiplication. (In linear algebra, probably you essentially proved this was a group without using the word group.) It is non-abelian if $n \geq 2$. (If you don't know matrix multiplication is non-commutative, pick two random

[3] Determining when two groups have the same structure is one of the basic problems in group theory. We remark it is a hard (as in research level) problem do determine exactly the number of different kinds of (the number of "isomorphism classes") of groups of a fixed order n. No exact formula is known (except for n of special type) and the number of distinct groups (up to isomorphism) of order n grows very quickly as the number of factors of n grows. (There is only type of group of order n when $n = p$ is prime, which is the isomorphism class of $\mathbb{Z}/p\mathbb{Z}$.)

2×2 matrices A, B and compute the products AB and BA.)

Example 3.3.8. Let $\mathrm{SL}(2,\mathbb{Z}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a,b,c,d \in \mathbb{Z},\ ad-bc=1 \right\}$. This is an infinite non-abelian group with usual matrix multiplication, and is an important group in number theory. To prove it is a group, the main point is to show that the matrix inverse of an element in $\mathrm{SL}(2,\mathbb{Z})$ is again in $\mathrm{SL}(2,\mathbb{Z})$. (Here it does not suffice to look at matrices with integer entries whose determinant is nonzero—you need that the determinant is a unit in $\mathbb{Z}$—e.g., $\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ has integer entries and nonzero determinant, but its inverse has fractional entries.)

Okay, so those were some examples. Basically, a group (in multiplicative notation) is a collection of objects that you can multiply and divide, and has "1." Recall we are interested in the invertible elements mod n, or equivalently, the invertible elements of $\mathbb{Z}/n\mathbb{Z}$:

$$(\mathbb{Z}/n\mathbb{Z})^\times = \{n\mathbb{Z} + a \in \mathbb{Z}/n\mathbb{Z} : a \text{ invertible mod } n\}.$$

More generally, for a ring R, we denote the set of **invertible** elements of R by $R^\times$, i.e.,

$$R^\times = \{a \in R : ab = 1 \text{ for some } b \in R\}.$$

Proposition 3.3.4. *Let R be a (commutative) ring. Then $R^\times$ is an abelian group. In particular, $(\mathbb{Z}/n\mathbb{Z})^\times$ is an abelian group for any n.*

Recall that for $a \in R$, a^{-1} denotes an inverse when it exists. Furthermore, by the same argument as for Exercise 1.2.4, an inverse is unique when it exists.

This result generalizes the earlier examples of $\mathbb{Q}^\times$, $\mathbb{R}^\times$ and $\mathbb{C}^\times$. Similarly, there is an analogue for non-commutative rings which generalizes the example of $\mathrm{GL}_n(\mathbb{R}) = M_n(\mathbb{R})^\times$.

Proof. Consider $a, b \in R^\times$, which have inverses a^{-1}, b^{-1}. Then $(ab)(b^{-1}a^{-1}) = a(bb^{-1})a^{-1} = aa^{-1} = 1$, so ab is also invertible, and thus $ab \in R^\times$. This means multiplication defines a binary operation on $R^\times$. Further, it is associative since multiplication on R is.

First note that $1 \in R^\times$, so $R^\times$ is non-empty and has a multiplicative identity. Next, if $a \in R^\times$, then there exists $a^{-1} \in R$ such that $aa^{-1} = a^{-1}a = 1$, so also $a^{-1} \in R^\times$, and thus $R^\times$ (essentially by definition) contains inverses. □

Proposition 3.3.5. *We have*

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

Hence $|(\mathbb{Z}/n\mathbb{Z})|^\times$ is the number of integers $1 \le a < n$ with $\gcd(a,n) = 1$.

Proof. Let $a \in \mathbb{Z}$. Note a is invertible mod n if and only if

$$ax + ny = 1 \tag{3.3.1}$$

has a solution for some $x, y \in \mathbb{Z}$. By the Euclidean algorithm (see Proposition 2.3.1), this happens if and only if $\gcd(a,n) = 1$. This proves the first statement, and the second statement follows immediately. □

Definition 3.3.6. *The function $\phi : \mathbb{N} \to \mathbb{N}$ given by $\phi(n) = |(\mathbb{Z}/n\mathbb{Z})^\times|$ (where we interpret $\phi(1) = 1$) is called the* **Euler phi** *or* **Euler totient** *function.*

Example 3.3.9. When $n = 2$, we have $(\mathbb{Z}/2\mathbb{Z})^\times$ consists of 1 element, $1 + 2\mathbb{Z}$. It is its own inverse. Thus $\phi(2) = 1$.

Example 3.3.10. When $n = 3$, we have $(\mathbb{Z}/3\mathbb{Z})^\times$ consists of 2 elements, $1+3\mathbb{Z}$ and $2+3\mathbb{Z}$. Since $1 \cdot 1 \equiv 1 \bmod 3$ and $2 \cdot 2 \equiv 1 \bmod 3$, we see they are each their own inverse. Thus $\phi(3) = 2$.

Recall that, to avoid the cumbersome notation $a + n\mathbb{Z}$, we often denote the elements of $\mathbb{Z}/n\mathbb{Z}$ using a set of representatives $\{0, 1, 2 \ldots, n-1\}$ from $\mathbb{Z}$, e.g., we will often write 2 instead of $2 + n\mathbb{Z}$. We hope this will not cause any confusion.

Example 3.3.11. For $n = 4$, a set of representatives for $(\mathbb{Z}/4\mathbb{Z})^\times$ is $\{1, 3\}$. Again, each element is its own inverse, and we see $\phi(4) = 2$.

Example 3.3.12. For $n = 5$, a set of representatives for $(\mathbb{Z}/5\mathbb{Z})^\times$ is $\{1, 2, 3, 4\}$, so $\phi(5) = 4$. We see $2 \cdot 3 \equiv 1 \bmod 5$ and $4^2 \equiv (-1)^2 \equiv 1 \bmod 5$, so 1 and 4 are their own inverses, while 2 and 3 are inverses of each other.

Exercise 3.3.2. For $6 \leq n \leq 10$, write down a set of representatives for $(\mathbb{Z}/n\mathbb{Z})^\times$, determine the inverse of each representative, and compute $\phi(n)$.

If (x, y) is a solution to (3.3.1), then x is an inverse to a mod n. Hence we can compute inverses of a mod n using the extended Euclidean algorithm/tableau method. This will be useful when n is very large, and is an important step in the RSA cryptosystem below.

Exercise 3.3.3. Use the extended Euclidean algorithm to find by hand an inverse to 37 mod 100. Check that your solution is indeed an inverse.

The above proposition readily gives:

Corollary 3.3.7. *For $n \geq 2$, we have $\phi(n) \leq n - 1$, with equality if and only if n is prime. Hence $\mathbb{Z}/n\mathbb{Z}$ is a field if and only if n is prime.*

Proof. Since there are only n elements of $\mathbb{Z}/n\mathbb{Z}$ and 0 is never invertible if $n \geq 2$, we immediately get $\phi(n) \leq n - 1$. If n is prime, then each $1 \leq a < n$ has $\gcd(a, n) = 1$, so $\phi(n) = n - 1$. If n is not prime, it has a nontrivial divisor $1 < m < n$. Then m is not invertible mod n by the above proposition, so $\phi(n) < n - 1$. This proves the first statement.

For the second, recall that $\mathbb{Z}/n\mathbb{Z}$ is a field if and only if each nonzero element is invertible, i.e., if and only if $\phi(n) = n - 1$. □

We've seen $\phi(n)$ is easy to compute when n is prime, and you might wonder about other values. Indeed, $\phi(n)$ is a basic function in number theory, and many elementary number theory courses derive a precise formula in terms of the prime factorization of n. We will just do a couple of special cases now, but see how to say something more general later.

Proposition 3.3.8. *For any prime p, $\phi(p^2) = p(p-1)$.*

Proof. We just need to count the numbers between 1 and $p^2 - 1$ which are relatively prime to p^2, i.e., relatively prime to p. Since p is prime, these are just the multiples of p up to $p^2 - 1$:

$$p, 2p, 3p, \ldots, (p-1)p,$$

of which there are $p - 1$. So we have $(p^2 - 1) - (p - 1) = p^2 - p = p(p - 1)$ numbers up to $p^2 - 1$ which are relatively prime to p. □

Exercise 3.3.4. Determine $\phi(p^n)$. Test your formula on small powers of 2 and 3.

The following situation will come up in RSA below:

Exercise 3.3.5. Prove $\phi(pq) = (p-1)(q-1)$ when p and q are distinct primes.

Exercise 3.3.6. Determine $\phi(60)$.

Exercise 3.3.7. Write $n = p_1^{e_1} \cdots p_r^{e_r}$ (the prime-power factorization). Conjecture a formula for $\phi(n)$ in terms of p_i's and e_i's, and provide some evidence for your conjecture.

3.4 Cosets and Lagrange's theorem

Definition 3.4.1. *Let $(G, \cdot)$ be a group. Let H be a subset of G. If $(H, \cdot)$ is also a group then H is called a* **subgroup** *of G. The* **(left) cosets** *of H in G are the subsets of G of the form*

$$g \cdot H = \{g \cdot h : h \in H\} \quad g \in G.$$

Just like the subring test from Lemma 1.2.6, we have the following subgroup test.

Lemma 3.4.2. *If G is a group and H is a nonempty subset of G, then H is a subgroup of G if and only if it is closed under multiplication ($h_1 h_2 \in H$ for $h_1, h_2 \in H$) and inversion ($h^{-1} \in H$ for $h \in H$).*

Proof. ($\Leftarrow$) Suppose H is closed under multiplication and inversion. Being closed under multiplication implies that the multiplication on G restricts to a well defined binary operation on H. Associativity holds because it does in G. If H is closed under inversion, then pick any $h \in H$ so $h^{-1} \in H$. (Here is where we need H nonempty.) By closure under multiplication $hh^{-1} = 1 \in H$. This takes care of all 3 properties required to be a group.

($\Rightarrow$) If H is a group, it is closed under multiplication and inverses by definition. □

Example 3.4.1. The set $n\mathbb{Z} \subseteq \mathbb{Z}$ consisting of multiples of n is a subgroup of $\mathbb{Z}$. To check this, obvserve $0 \in n\mathbb{Z}$ (so $n\mathbb{Z}$ is nonempty), the sum of two multiples of n is a multiple of n, and for any $kn \in n\mathbb{Z}$, $-kn \in n\mathbb{Z}$. Then the cosets of $n\mathbb{Z}$ in $\mathbb{Z}$ are the subsets of G of the form $a + n\mathbb{Z}$ for $a \in \mathbb{Z}$. In other words, the cosets of $n\mathbb{Z}$ in $\mathbb{Z}$ are precisely the congruences classes mod n.

Just as congruences mod n partition $\mathbb{Z}$ into n different classes, we will see in the proof of Lagrange's theorem below that the cosets partition G into a certain number of different classes (which we now call cosets). (In fact, in general we can view cosets as equivalence classes with the equivalence relation—see Exercise 3.4.4—but we will not emphasize this point of view in this course.) Let us first look at a few more examples.

Example 3.4.2. Let G be any group, and 1 the identity. Then it is easy to see $H = \{1\} \subseteq G$ is a subgroup, called the **trivial (sub)group**. For any $g \in G$, $g \cdot H = \{g\}$. Hence there are $|G|$ cosets of H in G, each consisting of a single element. This corresponds to the unique partition of G into $|G|$ singleton sets.

Example 3.4.3. Clearly $H = G$ is a subgroup of G. Then for any $g \in G$, $gH = H = G$ (a proof is contained in the proof of Lagrange's theorem below), so there is only one coset, $H = G$ itself. This corresponds to the "trivial partition" of G into one set, G.

Example 3.4.4. $\mu_2 = \{\pm 1\}$ is a subgroup of $\mu_4 = \{\pm 1, \pm i\}$. Note $1 \cdot \mu_2 = -1 \cdot \mu_2 = \mu_2$, and $i \cdot \mu_2 = -i \cdot \mu_2 = \{\pm i\}$. So there are two cosets of μ_2 in μ_4, and they give the following partition of μ_4:

$$\mu_4 = \mu_2 \sqcup i\mu_2 = \{1, -1\} \sqcup \{i, -i\}.$$

Exercise 3.4.1. Show that the only subgroups of μ_4 are $\mu_1 = \{1\}$, μ_2 and μ_4.

Example 3.4.5. Both μ_2, μ_3 are subgroups of $G = \mu_6 = \left\{\zeta_6^i : 0 \le i \le 5\right\} = \left\{\pm 1, \pm\zeta_6, \pm\zeta_6^2\right\}$.
First consider $H = \mu_2 = \{\pm 1\}$. Then the cosets are

$$1 \cdot \mu_2 = -1 \cdot \mu_2 = \{\pm 1\}, \quad \zeta_6 \cdot \mu_2 = \zeta_6^4 \cdot \mu_2 = \{\pm\zeta_6\}, \quad \zeta_6^2 \cdot \mu_2 = -\zeta_6^2 \cdot \mu_2 = \left\{\pm\zeta_6^2\right\}.$$

For $H = \mu_3 = \left\{1, \zeta_3, \zeta_3^2\right\} = \left\{1, \zeta_6^2, \zeta_6^4\right\}$, the cosets are

$$1 \cdot \mu_3 = \zeta_6^2 \cdot \mu_3 = \zeta_6^4 \cdot \mu_3 = \left\{1, \zeta_6^2, \zeta_6^4\right\}, \quad \zeta_6 \cdot \mu_3 = \zeta_6^3 \cdot \mu_3 = \zeta_6^5 \cdot \mu_3 = \left\{\zeta_6, \zeta_6^3, \zeta_6^5\right\}.$$

Exercise 3.4.2. Show that the only subgroups of μ_6 are μ_1, μ_2, μ_3 and μ_6.

You might have noticed that in the examples above that all cosets of H in G have the same size, and if $\{g_1, \dots, g_r\}$ is a coset, we can represent it as $g_i \cdot H$. All of this will fall out of the proof of our next result.

Proposition 3.4.3. (Lagrange's theorem) *Suppose H is a subgroup of a finite group G. Then there are $|G|/|H|$ distinct cosets of H in G, each of size $|H|$. In particular, $|H|$ divides $|G|$.*

Proof. First note that any the size of any coset gH is $|H|$: if $h, h' \in H$, then

$$gh = gh' \implies g^{-1}gh = g^{-1}gh' \implies h = h',$$

hence for a fixed g, all the products gh are distinct.

Now we claim that any two distinct g_1H and g_2H are disjoint. For if they intersect, then for some $h_1, h_2 \in H$, we have $g_1h_1 = g_2h_2$. We can write any $h \in H$ as $h_1h_1^{-1}h$, so

$$g_1h_1 = g_2h_2 \implies g_1h = g_1h_1h_1^{-1}h = g_2(h_2h_1^{-1}h) \in g_2H,$$

i.e., any element of g_1H must be inside g_2H also. But since the have the same size ($|H|$), we must have $g_1H = g_2H$, proving the claim.

Hence the cosets $\{gH\}$ partition G into disjoint subsets, all of size $|H|$. In particular there must be $|G|/|H|$ cosets, which proves Lagrange's theorem. □

Exercise 3.4.3. Let H be a subgroup of a finite group G, and $C = \{g_1, \dots, g_r\}$ a coset of H in G. Prove that, for $g \in G$, $g \cdot H = C$ if and only if $g \in C$.

Exercise 3.4.4. Let H be a subgroup of a finite group G. Define $g_1 \equiv g_2 \bmod H$ if $g_2^{-1}g_1 \in H$.
(i) Show $g_1 \equiv g_2 \bmod H$ if and only if $g_1H = g_2H$.
(ii) Prove this defines an equivalence relation on G, and that the equivalence classes are simply the cosets of H in G.

Exercise 3.4.5. Let $G = (\mathbb{Z}/8\mathbb{Z})^\times$, which we represent as $\{1, 3, 5, 7\}$.
(i) Write down the multiplication table for G.
(ii) Let $H = \{1, 7\}$. Show H is a subgroup of G.
(iii) Determine the cosets of H in G.

Exercise 3.4.6. Let $G = (\mathbb{Z}/7\mathbb{Z})^\times$. We represent the elements of G by $1, 2, \dots, 6$.
(i) Write down the multiplication table for G.
(ii) Let $H = \{1, 6\}$. Show H is a subgroup of G.
(iii) Determine the cosets of H in G.
(iv) Repeat (ii) and (iii) for the set $H = \{1, 2, 4\}$.

Exercise 3.4.7. Let $n > 2$. Recall D_{2n} is the symmetries of a regular n-gon P.
(i) Label the vertices of P by $1, 2, \dots n$. Use this to realize D_{2n} as a subgroup of the symmetric group S_n.
(ii) Show $D_6 = S_3$.
(iii) Determine the cosets of D_8 in S_4.

Lemma 3.4.4. *Let G be a finite group and $a \in G$.*
(i) There is some $n \in \mathbb{N}$ such that $a^n = 1$.
(ii) Take the smallest such n, called the **order** *of a. Then $C = \{a, a^2, a^3, \cdots, a^n\}$ is a subgroup of G of order n.*

Proof. (i) Since G is finite, and $a^k \in G$ for all $k \in \mathbb{N}$ there must be some $j, k \in \mathbb{N}$ with $j \neq k$ such that $a^j = a^k$. Assume $j < k$ and let $n = k - j$. Then $a^j a^n = a^j a^{k-j} = a^k = a^j$. Multiplying by $(a^j)^{-1}$, we see $a^n = 1$.

(ii) Let n be the order of a, i.e., $n \in \mathbb{N}$ is the smallest such that $a^n = 1$. Then the argument in (i) shows we can't have $a^j = a^k$ for $1 \leq j < k \leq n$—otherwise $a^{k-j} = 1$ but $k - j < n$. Hence C has precisely n elements.

By the lemma above, to check it is a subgroup it suffices to check closure under multiplication and inverses. Take any a^j and a^k in C (with $1 \leq j, k \leq n$). If $j + k \leq n$, $a^j a^k = a^{j+k} \in C$ trivially; if $j + k > n$, we see $a^j a^k = a^{j+k} = a^{j+k-n} a^n = a^{j+k-n} \in C$ since $1 \leq j + k - n \leq n$. Hence C is closed under multiplication.

Note since $a^n = 1$, $(a^n)^{-1} = 1 = a^n \in C$. For any $1 \leq j < n$, we have $1 \leq n - j < n$. Then since $a^j a^{n-j} = a^n$ □

The group C in this lemma is called the **cyclic subgroup generated by** a because it consists only of elements that are powers of a single element a. (It is called cyclic because these powers cyclically repeat: $a^n = 1, a^{n+1} = a^n a = a, a^{n+2} = a^n a^2 = a^2, \dots$.)

Exercise 3.4.8. Check that the powers of a cyclically repeat in this example.
(i) In $(\mathbb{Z}/7\mathbb{Z})^\times$), compute 3^k for $1 \leq k \leq 10$.
(ii) What is the cyclic subgroup of $(\mathbb{Z}/7\mathbb{Z})^\times$ generated by 3? What about generated by 2?

Proposition 3.4.5. *Let G be a finite group of order n. Then, for any $a \in G$, $a^n = 1$.*

Proof. Say m is the order of a in G. Then a generates a cyclic subgroup H of G of order m, by the previous lemma. Now by Lagrange's theorem, $m|n$, say $n = km$. Then

$$a^n = a^{km} = (a^m)^k = 1^k = 1.$$

□

The proof is essentially summarized in the following phrase: the order of any element of G divides the order of G.

Corollary 3.4.6. (Fermat's little theorem) *If p is prime and $a \not\equiv 0 \bmod p$, then $a^{p-1} \equiv 1 \bmod p$.*

Proof. Apply the previous proposition to $G = (\mathbb{Z}/p\mathbb{Z})^\times$, which has order $p-1$. □

Corollary 3.4.7. (Formula for inverses mod p**)** *Suppose* $\gcd(a,p) = 1$. *Then the inverse* a^{-1} *of* a *mod* p *is given by* $a^{-1} \equiv a^{p-2} \bmod p$.

Proof. Note $a^{-1}a \equiv a^{p-2}a \equiv a^{p-1} \equiv 1 \bmod p$. □

This gives a quick way to compute inverses mod p, using what is what is called the method of **repeated squaring**. We just illustrate this procedure with an example.

Example 3.4.6. Let's compute the inverse of 3 mod 19. By the above corollary, we have $3^{-1} \equiv 3^{17} \bmod 19$. We first use repeated squaring to compute the $3^{2^j} \bmod 19$ for $2^j \leq 17$:

$$\begin{aligned} 3^2 &\equiv 9 \bmod 19 \\ 3^4 \equiv (3^2)^2 &\equiv 9^2 \equiv 81 \equiv 5 \bmod 19 \\ 3^8 \equiv (3^4)^2 &\equiv 5^2 \equiv 25 \equiv 6 \bmod 19 \\ 3^{16} \equiv (3^8)^2 &\equiv 6^2 \equiv 36 \equiv 17 \bmod 19. \end{aligned}$$

Then we write 17 as a sum of powers of 2: $17 = 16 + 1$, and use this to compute

$$3^{-1} \equiv 3^{17} \equiv 3^{16+1} \equiv 3^{16} \cdot 3^1 \equiv 17 \cdot 3 \equiv (-2) \cdot 3 \equiv -6 \equiv 13 \bmod 19.$$

We can check this is indeed the inverse: $3 \cdot 13 \equiv 39 \equiv 1 \bmod 19$.

We also noted we can compute inverses mod p (in fact, mod n for any n) via the extended Euclidean algorithm in the last section. While that method is quite fast (and often faster than the above method), it is often useful in theory to have a *formula* rather than just an *algorithm*. On the other hand, it is sometimes more useful to have an algorithm than a formula, and we will see both Euler's theorem and the extended Euclidean algorithm being used in (different parts of) RSA in the next section.

Exercise 3.4.9. Use the formula $a^{-1} \equiv a^{p-2} \bmod p$ with repeated squaring to compute by hand the inverse of 5 mod 23. Check your answer is correct.

In fact, we will want to use the following generalization of Fermat's little theorem.

Corollary 3.4.8. (Euler's theorem) *For any invertible* $a \bmod n$, *we have* $a^{\phi(n)} \equiv 1 \bmod n$.

Proof. Apply the above proposition to $G = (\mathbb{Z}/n\mathbb{Z})^\times$, which, by definition of the totient function, has order $\phi(n)$. □

Exercise 3.4.10. Use Euler's theorem and repeated squaring to compute by hand $3^{-1} \bmod 14$.

As an addendum, we say a little more about cyclicity and orders. We say a finite group G is **cyclic** if there exists $g \in G$ such that the order of g is the order of G. Note for such a g, then the cyclic subgroup of G generated by g has order $|G|$, and so must be all of G. Note $(\mathbb{Z}/n\mathbb{Z}, +)$ and μ_n are cyclic groups of order n, and we can take for generators 1 and ζ_n, respectively. (By a **generator** of a cyclic group G, we mean any element of g which cyclically generates G, i.e., any element of order $|G|$.)

Exercise 3.4.11. For $2 \leq n \leq 10$, determine if $(\mathbb{Z}/n\mathbb{Z})^\times$ is cyclic or not. When it is cyclic, list all of the generators.

3.5 RSA

Beyond the uses of very elementary arithmetic, number theory has long been regarded largely as a purely theoretical study, with little practical applications, particularly when compared with fields like calculus and differential equations, which have many more obvious connections with real world applications. Things have greatly changed in our modern information age, and now aspects of number theory and algebra that were long considered purely theoretical have found many applications to computer science and information theory (so by extension are of interest to computer and electrical engineers as well). Two of the main sources of applications are cryptography and error-correcting codes, which can be bundled together under the name of coding theory.

Cryptography, the more famous of the two, is about keeping information secret from intruders, whereas error-correcting codes are about the opposite situation: how to send and receive information across a noisy channel (e.g., communicating with satellites). Both of these subjects are now a fundamental part of modern life, with most people not realizing what they are doing for us "under the hood." Essentially, any time you use a modern electronic device, you're relying on coding theory from things to making secure purchases online (or credit card purchases in store) and keeping other people from logging into your accounts (both cryptography), to having any sort of network reliability on a mobile device and not losing information on your hard drive anytime a butterfly flaps its wings (both error-correcting codes).

In this section, we'll just explain one beautiful and very practical application of number theory to cryptography: the RSA cryptosystem (the main idea, without all of the implementation details). To put this into context, let us first just very briefly discuss some general cryptography. Here is the basic problem in cryptography, which involves 3 characters:

- Alice, our protagonist. She want to send Bob a secret message.
- Bob, his name is Bob.
- Eve, the specific antagonist, and general ne'er-do-well. She eavesdrops on communications between Alice and Bob.

Problem: How can Alice send Bob a message in such a way that only Bob will be able to read it?

Private-key cryptography

The classical approach to this is using what is known as **private-key cryptography**. In this, Alice and Bob agree upon a secret code, or cipher, in advance. This involves 3 things: (i) an encryption algorithm, (ii) a decryption algorithm, and (iii) a secret key. One of the simplest and oldest ciphers is the **(Caesar) shift cipher**. Let's assume messages just

consist of letters A, B, ..., Z. Our encryption will be to just to cyclically shift the letters by k "to the right". E.g., if $k = 1$, A will get encrypted as B, B as C, and so on, until Z which gets encrypted to A. If $k = 2$, A gets encrypted to C, B to D, and so on until Y to A, and Z to B. To decrypt, you simply cyclically shift the letters "to the left" by k. Both encryption and decryption, besides requiring that we know we are using a Caesar shift, require knowing the number $0 \leq k \leq 25$ (note: $k = 0$ is not so great!), which we call the key for this cryptosystem.

For instance, if Alice wants to send Bob the message EVESUCKS, they might agree on a shift cipher with $k = 3$ in advance[4], so Alice would encrypt letter by letter, send Bob the message

HYHVXFNV

which Bob easily decrypts, and if Eve sees the message along the way, it looks like nonsense to her. However, if she knows or guesses that they are using a shift cipher, and really wants to decode it, she can try all possible values for k, and notice that decrypting with $k = 3$ gives the only message that makes sense, and figure out the message. Of course, since the time of Caesar, ciphers have gotten incredibly more complex, and for good cryptosystems are nigh impossible to crack even if you know the algorithm (but not the key) being used.[5]

Public-key cryptography

The main problem with private-key cryptography, is that both Alice and Bob need to know the key without Eve knowing. This is fine if Alice and Bob can meet in advance in private to decide upon a key, but if they can't, or if they need to choose a new key, this is going to be quite difficult. In the 1970's, cryptographers were thinking about a way around this issue of making Alice and Bob agree on a key in advance, which led to what is now called **public-key cryptography**. The basic idea is the following: Bob makes generates a two keys: a **public key** e for encryption and a **private key** d for decryption. The public key e he announces to the world, and Alice can use e to encrypt her message, and send it to Bob. Then Bob, and only Bob, can decrypt the message using d, as only he knows d.

This idea requires two things. First, a cryptosystem where the encryption and decryption keys are different. E.g., if d and e are inverses in a ring R, encryption of a ring element $m \in R$ could be multiplication by e, giving the encrypted message $x = em$, and decryption could be multiplication by d: $dx = (de)m = m$. Second, since everyone knows the public key e, it should be hard to determine the decryption key d from just only e, but it should be easy for Bob to generate a pair of keys (e, d). Note that in our toy example of multiplication by e and d in R, at least in the case $R = \mathbb{Z}/n\mathbb{Z}$, it is easy to compute d from e via Euler's theorem or the extended Euclidean algorithm, so this would not make a good public-key cryptosystem. Note it's not at all obvious that a cryptosystem is possible, and for awhile cryptographers weren't sure if it was.

[4] Julius Caesar reportedly used this shift with $k = 3$ to communicate with his generals.

[5] Typically, the weakest link in computer security is not the cryptosystem. Usually in hacking/data breach scandals, hackers are exploiting people not following good security protocols, rather than "cracking cryptosystems." E.g., people have easily guessable password, sensitive information is stored unencrypted, account number printouts are just found in a bank dumpster, a company allows someone to reset your password without really proving they are you, you download a virus that logs all your keystrokes, etc.

In 1978, Rivest, Shamir and Adleman published a paper with such a cryptosystem, now known as **RSA**, whose security is based upon the belief that factoring integers is hard. RSA is now widely used, and you are probably using RSA anytime you do something securely online. For instance, anytime you use an https website (e.g., any secure login webpage, credit card payment page, etc), both your web browser and the server are using RSA. When the server sends you information, they encrypt it with RSA using your browser's public key, and when you send information back you encrypt it with the server's public key.

Here is the basic algorithm, with explanations to follow:

(1) Bob chooses 2 large primes p and q, sets $n = pq$, and chooses some $1 < d, e < \phi(n)$ such that $de \equiv 1 \bmod \phi(n)$. Then Bob posts (e, n) as the public key, keeping, p, q and d secret, with d being the private key.

(2) Alice has a message m, which she represents as an integer $< n$. (If the message m is too long, she can break it up into pieces and encrypt each piece separately.) She encrypts it with Bob's public key as $x \equiv m^e \bmod n$, and sends the cipher text x to Bob.

(3) Bob decrypts the message x by computing $x^d \equiv m \bmod n$.

The first step, called key generation, only needs to be done once to initialize the process, and then Alice can send Bob as many encrypted messages as she wants with using Bob's fixed public key. The Prime Number Theorem, about distribution of primes, says that it's not too hard to find big primes. Basically, just choose a really big random number (say around 1,000 bits, or around 200 digits) and test nearby numbers to see if they are prime. What's important here are two things (i) there are primality tests which are fast (they don't rely on factoring integers)[6], and (ii) the Prime Number Theorem says you only need to try around $\log m$ numbers to find a prime near some big number m. Do this twice, once to find p, and once to find q. Then, calculation of $n = pq$ is not hard. Then Bob can just randomly choose $1 < d < \phi(n)$, and with probability near 1. Since we know $n = pq$, we know $\phi(n) = (p-1)(q-1)$ by Exercise 3.3.5, so we can quickly compute $\phi(n)$ also. Then we can quickly invert d mod $\phi(n)$ with the extended Euclidean algorithm to get $1 < e < \phi(n)$ such that $de \equiv 1 \bmod \phi(n)$.[7] These are all the calculations needed for Step 1.

Example 3.5.1. To work with a small example, say Bob wants to take p, q to each be 3 digits long. (I did the following calculations in the Sage mathematical software package.) We randomly generate a couple numbers between 100 and 1000. I got 582 and 959. Starting

[6]Here a fast probabilistic primality test to see if for some integer m is most likely prime: First, you can use divisibility tests to quickly check for divisibility of m by small primes. If m is divisible by some small prime, we know m is not prime. This rules out most numbers quickly: e.g., $1/2$ of numbers are divisible by 2, $1/2 + 1/3 - 1/6 = 2/3$ of numbers are divisible by 2 or 3, and so on. If m is not divisible by a small prime, we can take a random number a less than m, and compute by repeated squaring a^{m-1} mod m—if m is prime, this is $\equiv a$ mod m by Fermat's little theorem. But if m is not prime, it turns out it's very unlikely that $a^{m-1} \equiv a$ mod m (though it happens occasionally). So if $a^{m-1} \equiv a$ mod m, we conclude m is probably prime (and we can try this for a few values of a if we like). Otherwise, we showed m is not prime.

[7]Here it's *much* better to use the extended Euclidean algorithm as opposed to Euler's theorem to compute e—if we tried to use Euler's theorem, we'd need to compute $\phi(\phi(n))$, which essentially requires factoring $\phi(n)$, which may be infeasible as we're working with very big numbers in practice.

with 582, I test successive numbers for primality, and I get $p = 587$ is prime, and similarly $q = 967$ is prime. So Bob computes $n = pq = 567629$ and $\phi(n) = (p-1)(q-1) = 566076$. Now we randomly take a number between 1 and 566076, say 154951. Using the Euclidean algorithm, we find 154951 is not invertible mod $\phi(n)$—their gcd is 23. Testing the next few numbers, we see $d = 154955$ is invertible mod $\phi(n)$, and its inverse is $e = 402575$. So Bob publishes e and n as his public key.

Exercise 3.5.1. Say Bob takes $p = 7$, $q = 11$ and $d = 17$. Determine Bob's public key.

Alice's part in this is easy. She has some message m, which she can realize as a number in some standard way. In practice this is a file, which is written in binary, that you can break up into bite-size blocks and encrypt separately. What's important is that $m < n$. Then, since she (and Eve and everyone else) knows n and e, she can compute $x \equiv m^e \bmod n$ (interpreted as a number between 0 and $n-1$) quickly using repeated squaring, as discussed in the previous section.

Example 3.5.2. Continue with the set up from the previous example.

Let's say Alice, again wants to send the message EVESUCKS to Bob. We can convert this to a number as follows: realize each letter as a number between 0 and 25 in the obvious way (A=0, B=1, ..., Z=25). (If you wanted to, you could include punctuation, and what not into your conversion scheme, say using the ASCII code which represents each character as a number between 0 and 255. Or just view a computer file, which is stored as a string of 1's and 0's, as a number in binary.) So we can think of EVESUCKS as representing a base 26 number of length 8. Since $26^3 < n$, we can break this up into blocks of length 3 as EVE, SUC, and KSZ. (Here I needed to pad the last block with some symbol such as Z. In practice, you can use a special character just for padding.) Let's just do the first block, EVE. Since E corresponds to 4 and V corresponds 21, EVE represents the base 26 form of the number $m = 4 * 26^2 + 21 * 26 + 4 * 1 = 3254$ in decimal. Then Alice computes the cipher text (encrypted message) using repeated squaring as

$$x = (m^e \bmod n) = 391820.$$

Exercise 3.5.2. Using Bob's public key from the previous exercise, encrypt the message $m = 15$.

Just like the previous step, Bob's decryption is easy. Since $de = k\phi(n) + 1$ for some k, we have

$$x^d \equiv (m^e)^d \equiv m^{k\phi(n)+1} \equiv (m^{\phi}(n))^k m \equiv m \bmod n,$$

using Euler's theorem at the last step. (The original paper of RSA gave a slightly different proof that $x^d \equiv m \bmod n$ using Fermat's little theorem.) So when Bob computes $x^d \bmod n$ (again by repeated squaring), he recovers m.

Example 3.5.3. Continue with the setup from the previous two examples.

Bob receives the message $x = 320576$ as the first encrypted block of the message. He computes

$$(x^d \bmod n) = 3254 = m.$$

Now, Bob can convert this back to the first block of the plain text (unencrypted) message EVE.

We remark that, in practice, there are various standard protocols to automate the way to messages (or files) are converted into blocks of numbers, but our goal is not to get into the technical aspects of implementation on a computer, just the main idea of RSA.

Exercise 3.5.3. Perform Bob's decryption of the Alice's message from the previous exercise.

Now, why is this algorithm (believed to be) secure? Well, let's suppose Eve intercepts the cipher text x. To decrypt, she needs to know what to exponentiate it by (mod n) to get back to m. That is, she needs some d' such that $m^{d'e} \equiv m \bmod n$. Euler's theorem says this will be true if d' is an inverse of e mod $\phi(n)$. While there are a few choices of messages where other d''s can work (e.g., if $m = 1$, then $m^{d'e} \equiv m$ for any d'), for almost all messages m you really need d' to be an inverse of e mod $\phi(n)$. By the extended Euclidean algorithm, Eve can compute the inverse of e mod $\phi(n)$ to get Bob's decryption key d if she knows $\phi(n)$. But the point is that there are no known fast ways to compute $\phi(n) = (p-1)(q-1)$ without knowing p and q, and there are no known fast ways to factor n to get p and q (remember p, q, n are very large — hundreds of digits long).[8]

We remark the actual implementation of RSA involves a little more to avoid encountering special situations where the message can be easy to decrypt (e.g., if $m = 1$, or d or e is too small). Also, since the encryption and decryption process in RSA is slower than many private-key methods such as AES (the current government standard), sometimes RSA is used to exchange a private-key when sending large amounts of encrypted data.

Moreover, RSA can be used for **message authentication**. What's to prevent Eve from intercepting x and sending Bob a different fake message m' (which she can also encrypt with Bob's public key)? Well, if Alice wants to authenticate her message, she can add a **digital signature**. Basically the idea is to run RSA in reverse. First, she generates her own public key (n_A, e_A) and private key d_A. She can encrypt her message m using her *private* key d_A as $s \equiv m^{d_A} \bmod n_A$. This is her signature, and she can send both m and s. Then anyone in the world can check that s decrypts to m using Alices public key: $s^{e_A} \equiv m^{d_A e_A} \equiv m \bmod n_A$. Since no one else could generate s from m without knowing d_A, this proves m is from Alice. So Alice could instead of just sending Bob m, she could send him the pair (m, s) to prove a message wants to come from her, and if she doesn't want Eve to be able to read the message, she can first encrypt both m and s using Bob's public key. (She needs to encrypt s also, otherwise Eve could decode s to get m from Alice's public key.)

[8] As of 2024, the largest "RSA number" ($n = pq$, p, q both large) factored was 250 digits, and this required about 2700 CPU years and was completed in 2020. In applications, n is about 3–4 times larger than this. Since computational difficulty grows roughly exponentially, RSA is believed to extremely secure, unless quantum computing becomes feasible.

If you're interested in learning cryptography, there are many good references out there. One possibility is William Stein's *Elementary Number Theory* book mentioned in the introduction. We also have a course here in the math department, *Applied Modern Algebra*, whose actual content varies according to the instructor, but it is usually largely cryptography. (The last time I taught it it was 75% cryptography and 25% error-correcting codes, but another faculty who teaches it often makes it 100% cryptography.)

Chapter 4

Sums of Squares

In this chapter, will get our first major theorem about Diophantine equations: Fermat's determination of when a number is a sum of two squares. This will put together much of what we have learned in previous chapters, which were in some sense preliminaries to this and later theorems. The proof will use unique factorization in $\mathbb{Z}[i]$, norms, and modular arithmetic.

Then we will consider some related questions also studied by Fermat: when is a number of the form $x^2 + dy^2$, e.g., $x^2 + 2y^2$ or $x^2 + 3y^2$. For this, we will need two major theorems in elementary number theory: the Chinese Remainder Theorem and Quadratic Reciprocity. (Really, the main use of the Chinese Remainder Theorem is to prove Quadratic Reciprocity, which is one of Gauss's major contributions to number theory.) This will allow us to say some things about numbers of the form $x^2 + dy^2$, but a complete answer is not so easy.

Finally, we will briefly discuss the problems of when a number is a sum of three or four squares, which were answered by Gauss and Lagrange.

4.1 Sums of Two Squares

In this section, we will give a complete answer to the question: *what numbers are sums of two squares?* i.e., for what $n \in N$ does

$$x^2 + y^2 = n \tag{4.1.1}$$

have a solution $(x, y) \in \mathbb{Z} \times \mathbb{Z}$. The answer was known to Fermat, though our approach, which comes via unique factorization in $\mathbb{Z}[i]$, did not come until after Gauss. Recall the following, which we will repeatedly use:

Fact 4.1.1. *An integer n is a sum of two squares if and only if $n = x^2+y^2 = (x+yi)(x-yi) = N(x + yi)$ is a norm from $\mathbb{Z}[i]$.*

The fact above immediately yields the

Proposition 4.1.2. (Composition law) *If m and n are sums of two squares, so is mn.*

Proof. If m and n are sums of two squares, then $m = N(\alpha)$ and $n = N(\beta)$ for some $\alpha, \beta \in \mathbb{Z}[i]$. Then $mn = N(\alpha)N(\beta) = N(\alpha\beta)$ by multiplicativity of the norm, when mn is also a norm, i.e., a sum of two squares. □

Exercise 4.1.1. If $m = x^2+y^2$ and $n = z^2+w^2$, explicitly find u, v (in terms of x, y, z, w) such that $mn = u^2 + v^2$.

We also recall from Proposition 3.2.4 that (4.1.1) does not have a solution if $n \equiv 3 \bmod 4$.

The composition law suggests that the essential case of (4.1.1) is when $n = p$ prime. Indeed this is true, and we will first treat $n = p$. Since $2 = 1^2 + 1^2$, it suffices to answer this for $p \equiv 1 \bmod 4$. Here, a couple of auxiliary results will be useful.

Proposition 4.1.3. (Wilson's theorem) *Let p be a prime. Then $(p-1)! \equiv -1 \bmod p$.*

Proof. This is clear when $p = 2$, so assume p is odd. Recall each $1 \le a \le p-1$ is invertible mod p. Further a is its own inverse mod p if and only if $a^2 \equiv 1 \bmod p$, i.e., $p \mid (a^2-1)$. By the prime divisor property, this happens exactly when $p \mid (a-1)$ or $p \mid (a+1)$, but the bounds on a imply this happens if and only if $a = 1$ or $a = p-1$. So

$$(p-1)! \equiv \prod_{a=1}^{p-1} a \equiv 1 \cdot (p-1) \prod_{a=2}^{p-2} a \bmod p.$$

Now in the latter product (which must consist of an even number of terms, 0 if $p = 3$), each $2 \le a \le p-2$ has an inverse mod p which is some $2 \le a^{-1} \le p-2$ with $a^{-1} \ne a$. By uniqueness of inverses, we can group this latter product in to pairs of the form (aa^{-1}), whence the latter product is 1 mod p, so $(p-1)! \equiv -1 \bmod p$. □

Lemma 4.1.4. (Lagrange's lemma) *Let $p \equiv 1 \bmod 4$. Then -1 is a square mod p, i.e., there exists $m \in \mathbb{Z}$ such that $p \mid (m^2+1)$.*

Proof. First note that -1 is a square mod p means there exists $m \in \mathbb{Z}$ such that $m^2 \equiv -1 \bmod p$, i.e., $m^2 + 1 \equiv 0 \bmod p$, so indeed the two assertions in the statement of the lemma are equivalent.

Write $p = 4k+1$ for some $k \in \mathbb{N}$. By Wilson's theorem,

$$(4k)! \equiv -1 \bmod p.$$

On the other hand,

$$\begin{aligned}(4k)! \equiv (2k)! \times (2k+1)(2k+2)\cdots(4k) &\equiv (2k)! \times (-2k)(-2k+1)\cdots(-1) \\ &\equiv (2k)!(-1)^{2k}(2k)! \equiv ((2k)!)^2 \bmod p,\end{aligned}$$

hence -1 is a square mod p. □

Theorem 4.1.5 (Fermat)**.** *Let p be prime. Then $p = x^2 + y^2$ for some $x, y \in \mathbb{Z}$ if and only if $p = 2$ or $p \equiv 1 \bmod 4$.*

Proof. As remarked above, we already know $2 = 1^2 + 1^2$ and p is not a sum of 2 squares if $p \equiv 3 \bmod 4$ by Proposition 3.2.4. Thus it suffices to assume $p \equiv 1 \bmod 4$ and show p is a sum of 2 squares, i.e., show p is a norm from $\mathbb{Z}[i]$.

Note that if p is a reducible element of $\mathbb{Z}[i]$, we can write $p = ab$ for some $a, b \in \mathbb{Z}[i]$ with $N(a), N(b) > 1$. Since $N(a)N(b) = N(p) = p^2$, this means p is a norm from $\mathbb{Z}[i]$.

Suppose p is not a norm from $\mathbb{Z}[i]$. By the last paragraph, this means p is an irreducible element of $\mathbb{Z}[i]$. By unique factorization for $\mathbb{Z}[i]$, this means p is a prime of $\mathbb{Z}[i]$ (Theorem 2.5.1). Now by Lagrange's lemma, there exists $m \in \mathbb{Z}$ such that $p \mid (m^2+1) = (m+i)(m-i)$. Since p is prime in $\mathbb{Z}[i]$, this means $p \mid (m+i)$or $p \mid (m-i)$. But this is impossible as $\frac{m}{p} \pm \frac{i}{p} \notin \mathbb{Z}[i]$, contradicting the hypothesis that p is a norm from $\mathbb{Z}[i]$. □

Exercise 4.1.2. Let p be a prime of $\mathbb{N}$. Show p is a prime of $\mathbb{Z}[i]$ if and only if $p \equiv 3 \bmod 4$.

Exercise 4.1.3. Let p be a prime of $\mathbb{N}$. If $p = 2$ or $p \equiv 1 \bmod 4$, show that the irreducible factorization of p in $\mathbb{Z}[i]$ is of the form $p = \pi\overline{\pi}$, where π is any element of $\mathbb{Z}[i]$ of norm p.

Exercise 4.1.4. Show that the primes (i.e., irreducibles) of $\mathbb{Z}[i]$ are precisely the elements of the form (i) up where $u \in \{\pm 1, \pm i\}$ and $p \equiv 3 \bmod 4$ is a prime of $\mathbb{N}$, or (ii) an element of $\mathbb{Z}[i]$ of norm 2 or some prime $p \equiv 1 \bmod 4$. Further, show if π is an irreducible of the second type, then $u\pi \notin \mathbb{Z}$ for any unit u.

The next exercise is about counting the number of solutions to our favorite Diophantine equation.

Exercise 4.1.5. Let p be a prime of $\mathbb{N}$.
(i) Determine the number of irreducible elements of norm p in $\mathbb{Z}[i]$.
(ii) Deduce that for $p = 2$, there are exactly 4 solutions to $x^2 + y^2 = p$ with $x, y \in \mathbb{Z}$, and exactly 1 solution with $x, y \in \mathbb{N}$.
(iii) Deduce that for $p \equiv 1 \bmod 4$, there are exactly 8 solutions to $x^2 + y^2 = p$ with $x, y \in \mathbb{Z}$, and exactly 2 solutions with $x, y \in \mathbb{N}$.

Theorem 4.1.6. (Fermat's two square theorem) *Let $n \in \mathbb{N}$. Then n is a sum of two squares, i.e., $n = x^2 + y^2$ for some $x, y \in \mathbb{Z}$, if and only if each prime which is $3 \bmod 4$ appears to an even power in the prime-power factorization of n.*

Proof. Let us write the prime-power factorization of n as

$$n = \prod p_i^{e_i} \prod q_j^{f_j}$$

where each $p_i \equiv 3 \bmod 4$ and each q_j is 2 or $1 \bmod 4$.

($\Leftarrow$) First suppose the latter condition is satisfied, i.e., each e_i is even. Then $\prod p_i^{e_i}$ is a square, whence a sum of two squares. Also, by Theorem 4.1.5, we know each q_j is a sum of two squares. Then by the composition law, n is a sum of two squares.

($\Rightarrow$) To prove the converse direction, we essentially want a kind of converse to the composition law—that if rs is a sum of two squares then r and s must each be sums of two squares. This is obviously not true if $r = s$, but it turns out to be true if r and s are relatively prime, which the following argument shows. (See corollary below.)

Suppose n is a sum of two squares, i.e., $n = N(\alpha)$ for some $\alpha \in \mathbb{Z}[i]$. By the above exercises, each p_i is irreducible in $\mathbb{Z}[i]$ and an irreducible factorization of any q_j looks like $q_j = \pi_j \overline{\pi}_j$ where π_j is an element of norm q_j in $\mathbb{Z}[i]$. So an irreducible factorization of n in $\mathbb{Z}[i]$ looks like

$$n = \prod p_i^{e_i} \prod \pi_j^{f_j} \prod \overline{\pi}_j^{f_j}.$$

Now write an irreducible factorization of $\alpha \in \mathbb{Z}[i]$ as

$$\alpha = u \prod r_i^{h_i} \prod \phi_j^{k_j},$$

where u is a unit and, by Exercise 4.1.4, we may assume each r_i is a prime of $\mathbb{N}$ with $r_i \equiv 3 \bmod 4$ and each ϕ_j is an element of $\mathbb{Z}[i]$ of s_j, where s_j is a prime of $\mathbb{N}$ which is 2 or $1 \bmod 4$. Then, by multiplicativity of the norm,

$$n = N(\alpha) = N(u) \prod N(r_i)^{h_i} \prod N(\phi_j)^{k_j} = \prod r_i^{2h_i} \prod s_j^{k_j}.$$

Now, by unique factorization in $\mathbb{Z}$, we have up to reordering each $r_i = p_i$, $2h_i = e_i$, $s_j = q_j$ and $k_j = f_j$. Hence each e_i is even, which is precisely the latter condition in the theorem. □

The following structural result (a converse to the composition law) follows directly from the theorem:

Corollary 4.1.7. *Let $m, n \in \mathbb{N}$ with $\gcd(m, n) = 1$. Then mn is a sum of two squares if and only if both m and n are.*

Exercise 4.1.6. Suppose $p_1, \dots, p_r$ are distinct primes which are all $1 \bmod 4$. Determine the number of solutions to $x^2 + y^2 = p_1 \cdots p_r$.

Exercise 4.1.7. Suppose $p \equiv 3 \bmod 4$ and $q \equiv 1 \bmod 4$ are primes. Determine the number of solutions to $x^2 + y^2 = p^e q^f$ for $e, f \in \mathbb{N}$.

4.2 Pythagorean Triples

We can also apply the ideas from the last section to the determination of **Pythagorean triples** (x, y, z), i.e., positive integer solutions[1] to

$$x^2 + y^2 = z^2. \tag{4.2.1}$$

We say a Pythagorean triple $(x, y, z) \in \mathbb{N}^3$ is **primitive** if $\gcd(x, y) = 1$. If (x', y', z') is a triple and $\lambda = \gcd(x', y')$, then also $\lambda \mid z'$ and we can write $(x', y', z') = (\lambda x, \lambda y, \lambda z)$. Moreover (x', y', z') is a Pythagorean triple if and only if (x, y, z) is a primitive Pythagorean triple, so it suffices to determine primitive Pythagorean triples.

[1] We could also look at integer solutions to (4.2.1), but if (x, y, z) is a solution, then so is $(\pm x, \pm y, \pm z)$, and if one of x, y, z is 0, then the solutions are trivial—e.g., all integer solutions with $y = 0$ are $(x, 0, \pm x)$ for $x \in \mathbb{Z}$. Hence we get all (algebraically) interesting solutions to the Pythagorean equation by assuming $x, y, z > 0$, where this equation has the usual interpretation in terms of right-angled triangles.

Lemma 4.2.1. *Suppose (x, y, z) is a primitive Pythagorean triple. Then $x + yi$ and $x - yi$ are relatively prime in $\mathbb{Z}[i]$, i.e., they have no common prime divisors in $\mathbb{Z}[i]$.*

Proof. Suppose instead, $x + yi$ and $x - yi$ have a common prime divisor $\pi \in \mathbb{Z}[i]$. Then π divides their sum $2x$ and their difference $2yi$. Note if $\pi \mid x$ and $\pi \mid y$ then $1 < N(\pi) \mid N(x) = x^2$ and $1 < N(\pi) \mid N(y) = y^2$, but this is impossible if $\gcd(x, y) = 1$. Hence, $\pi \mid 2$, i.e., $\pi = \pm(1 \pm i)$. Then

$$N(\pi) = \pi\overline{\pi} = 2 \mid (x + yi)(x - yi) = x^2 + y^2 = z^2.$$

This means z is even, so $x^2 + y^2 \equiv z^2 \equiv 0 \bmod 4$, which implies x and y are also both even (use the same argument as in Proposition 3.2.5), contradicting primitivity. □

Lemma 4.2.2. *Suppose $\alpha, \beta \in \mathbb{Z}[i]$ are relatively prime. If $\alpha\beta = \gamma^2$ is a square in $\mathbb{Z}[i]$, then $u\alpha$ and $u^{-1}\beta$ are squares for some unit u of $\mathbb{Z}[i]$.*

Proof. Note that this is trivial if γ is a unit (and vacuous if $\gamma = 0$). So assume $\alpha\beta$ is the square of some $\gamma \in \mathbb{Z}[i]$, where γ is a non-zero non-unit. Then γ has a prime factorization in $\mathbb{Z}[i]$:

$$\gamma = \prod \pi_i^{e_i}.$$

Thus the prime factorization of $\alpha\beta$ is

$$\alpha\beta = \prod \pi_i^{2e_i}.$$

Up to a reordering of primes, we have

$$\alpha = u^{-1}\pi_1^{2e_1} \cdots \pi_j^{2e_j}$$

$$\beta = u\pi_{j+2}^{2e_{j+1}} \cdots \pi_k^{2e_k}$$

for some unit u. □

Exercise 4.2.1. Give an example of relatively prime non-units α, β in $\mathbb{Z}[i]$ such that $\alpha\beta$ is a square in $\mathbb{Z}[i]$, but α and β are not squares in $\mathbb{Z}[i]$.

Exercise 4.2.2. Show that if $u, v \in \mathbb{N}$ are relatively prime with $2 \mid uv$, then $(u^2 - v^2, 2uv, u^2 + v^2)$ is a primitive Pythagorean triple.

Proposition 4.2.3. *(x, y, z) is a primitive Pythagorean triple if and only if x and y are (in some order) $u^2 - v^2$ and $2uv$ for u, v relatively prime in $\mathbb{N}$ with $u > v$ and u, v not both odd. In this case, $z = u^2 + v^2$.*

Proof. ($\Leftarrow$) This is Exercise 4.2.2.

($\Rightarrow$) Suppose (x, y, z) is a primitive Pythagorean triple, so $x^2+y^2 = (x+yi)(x-yi) = z^2$. By the Lemma 4.2.1, $x + yi$ and $x - yi$ are relatively prime, and by Lemma 4.2.2 they are units times squares. In particular $x+yi = \pm\alpha^2$ or $x+yi = \pm i\alpha^2$ for some $\alpha \in \mathbb{Z}[i]$. Since -1 is a square in $\mathbb{Z}[i]$, we may absorb the possible minus sign into α and write either $x+yi = \alpha^2$ or $x + yi = i\alpha^2$.

Write $\alpha = u + vi$, and we get that either

$$x + yi = (u + vi)^2 = u^2 - v^2 + 2uvi$$

or

$$x + yi = i(u + vi)^2 = -2uv + (u^2 - v^2)i.$$

In the first case we have $x = u^2 - v^2$, $y = 2uv$. In the second, we may replace u by $-u$ to write $x = 2uv$, $y = u^2 - v^2$. It is easy to see the conditions $\gcd(u, v) = 1$, $u > v$ and u, v not both odd are necessary from the facts that $\gcd(x, y)$ and $x, y > 0$. (You will probably see this in the course of doing Exercise 4.2.2.)

In this setting, we have $z^2 = x^2+y^2 = N(x+yi) = N((u+vi)^2) = N(u+vi)^2 = (u^2+v^2)^2$, so $z = u^2 + v^2$. □

Corollary 4.2.4. *Let $p \in \mathbb{N}$ be prime. Then p occurs as the hypoteneuse of a right-angle triangle with integer length sides if and only if $p > 2$ is a sum of two squares, which is true if and only if $p \equiv 1 \bmod 4$.*

Proof. The second equivalence is Fermat's two square theorem, so it suffices to prove the first.

($\Rightarrow$) Suppose p is such a hypotenuse. Clearly $p \neq 2$. Now $x^2 + y^2 = p^2$. This implies $\gcd(x, y) = 1$. Hence by the proposition $p = u^2 + v^2$ for some u, v.

($\Leftarrow$) Suppose $p = u^2+v^2$ is odd. Then $u \neq v$ and u and v are not both odd. Furthermore, we may assume $u > v$. By the proposition $(u^2 - v^2, 2uv, p)$ is a primitive Pythagorean triple. □

Exercise 4.2.3. Let p, q be distinct primes. Determine when pq is the hypoteneuse of a right-angle triangle with integer length sides.

4.3 The Chinese Remainder Theorem

Recall that a key component of Fermat's two squares theorem was the determination of when -1 is a square mod p. To generalize Fermat's 2 squares theorem to other situations, e.g., what numbers (or primes) are of the form $x^2 + dy^2$, one is naturally led to the problem of what numbers are squares mod p.

This is addressed by Gauss's famous law of quadratic reciprocity, which Gauss called the "golden theorem." It first appeared in his *Disquisitiones Arithmeticae* (1804, but written in 1801 when he was 21). He thought it so important that he published 6 different proofs (now

there are at least 240 proofs!), and it is commonly regarded as the crown jewel of elementary number theory.

The proof we will give (in the next section) uses another famous result (much much older) from elementary number theory, the Chinese remainder theorem (CRT). This goes back over 1500 years ago to a book by Sun Tzu (no, not that Sun Tzu) from somewhere between the 3rd and 5th centuries. (Fun fact: now even in Chinese it's called (what translates to) the Chinese remainder theorem.) Another use of the CRT is to compute $\phi(n)$.

Theorem 4.3.1. (Chinese Remainder Theorem (CRT)) *Let $m, n \geq 2$ be relatively prime. Consider the map $\alpha : \mathbb{Z}/mn\mathbb{Z} \to (\mathbb{Z}/m\mathbb{Z}) \times (\mathbb{Z}/n\mathbb{Z})$ defined by sending any $a + mn\mathbb{Z}$ to $(a + m\mathbb{Z}, a + n\mathbb{Z})$ for any $a \in \mathbb{Z}$. Then α is a bijection, and moreover, restricted to $(\mathbb{Z}/mn\mathbb{Z})^\times$ gives a bijection of $(\mathbb{Z}/mn\mathbb{Z})^\times$ with $(\mathbb{Z}/m\mathbb{Z})^\times \times (\mathbb{Z}/n\mathbb{Z})^\times$.*[2]

Proof. First note that α is well defined, i.e., if $a \equiv b \bmod mn$, then $a + m\mathbb{Z} = b + m\mathbb{Z}$ and $a + n\mathbb{Z} = b + n\mathbb{Z}$, so $\alpha(a + mn\mathbb{Z})$ does not depend upon the choice the element a within a class $C = a + mn\mathbb{Z}$.

To show α is a bijection of $\mathbb{Z}/mn\mathbb{Z}$ with $\mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$, since both sets have size mn, it suffices to show it is an injection, i.e., it is one-to-one, i.e., no two elements of $\mathbb{Z}/mn\mathbb{Z}$ go to the same element of $\mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$ under α. Suppose $\alpha(a + mn\mathbb{Z}) = \alpha(b + mn\mathbb{Z})$. We may assume $0 \leq a \leq b < mn$. Then $a \equiv b \bmod m$ and $a \equiv b \bmod n$. Hence $b - a$ is divisible by both m and n, and thus by mn since m and n are relatively prime (here unique factorization is used too). But $0 \leq b - a < mn$, so this is only possible if $b - a = 0$, i.e., $a = b$, which proves α is one-to-one.

Last, we show α is a bijection of $(\mathbb{Z}/mn\mathbb{Z})^\times$ with $(\mathbb{Z}/m\mathbb{Z})^\times \times (\mathbb{Z}/n\mathbb{Z})^\times$. Recall that $a + mn\mathbb{Z} \in (\mathbb{Z}/mn\mathbb{Z})^\times$ if and only if $\gcd(a, mn) = 1$, i.e., if and only if $\gcd(a, m) = 1$ and $\gcd(a, n) = 1$. Hence for such an a, $\alpha(a + mn\mathbb{Z}) \in (\mathbb{Z}/m\mathbb{Z})^\times \times (\mathbb{Z}/n\mathbb{Z})^\times$. Conversely, given any element of $(\mathbb{Z}/m\mathbb{Z})^\times \times (\mathbb{Z}/n\mathbb{Z})^\times$, we can write this element in the form $(a+m\mathbb{Z}, a+n\mathbb{Z})$ for some $a \in \mathbb{Z}$ using the fact that α is a bijection of $\mathbb{Z}/mn\mathbb{Z}$ with $\mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$ (really, we only need that α is surjective, i.e. onto, for this). Similarly, for such a we have $\gcd(a, m) = \gcd(a, n) = 1$, which means $a + mn\mathbb{Z} \in (\mathbb{Z}/mn\mathbb{Z})^\times$. Now we have shown that α maps $(\mathbb{Z}/mm\mathbb{Z})^\times$ both into and onto $(\mathbb{Z}/m\mathbb{Z})^\times \times (\mathbb{Z}/n\mathbb{Z})^\times$. From the previous paragraph, we also know α restricted to invertible elements is an injection, so it must be a bijection. □

Corollary 4.3.2. *Let $m, n \geq 2$ be relatively prime. Then $\phi(mn) = \phi(m)\phi(n)$.*

Note this corollary gives $\phi(pq) = (p-1)(q-1)$ for distinct primes p, q as a special case, which was Exercise 3.3.5. Moreover, applying this corollary repeatedly gives us a formula for $\phi(n)$: if $n = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r}$, then

$$\phi(n) = \phi(p_1^{e_1})\phi(p_2^{e_2}) \cdots \phi(p_r^{e_r}). \tag{4.3.1}$$

(In a similar manner, we could state the CRT for $\mathbb{Z}/n_1 n_2 \cdots n_r\mathbb{Z}$ where the n_i's are relatively prime.) If each $e_i = 1$ (so n is square-free), then we just get $\phi(n) = (p_1-1)(p_2-1)\cdots(p_r-1)$. For arbitrary n, you can combine (4.3.1) with Exercise 3.3.4, to write down a similar formula $\phi(n)$ in terms of only the p_i's and e_i's, giving a definite answer to Exercise 3.3.7.

[2] For those who have had some algebra, in fact α is a ring isomorphism from $\mathbb{Z}/mn\mathbb{Z}$ to $(\mathbb{Z}/m\mathbb{Z}) \times (\mathbb{Z}/n\mathbb{Z})$ and restricts to a group isomorphism from $(\mathbb{Z}/mn\mathbb{Z})^\times$ to $(\mathbb{Z}/m\mathbb{Z})^\times \times (\mathbb{Z}/n\mathbb{Z})^\times$. In this way, the second statement (group isomorphism) follows from the first by restricting to the unit groups of the appropriate rings. The group isomorphism part (without using this terminology) is also Exercise 4.3.5.

Exercise 4.3.1. Use (4.3.1) to compute $\phi(60)$.

Exercise 4.3.2. If $n = p_1^{e_1} p_2^{e_2} \cdots p_r^{e_r}$, write an explicit formula for $\phi(n)$ in terms of only the p_i's and e_i's.

Exercise 4.3.3. How many numbers $1 \leq n \leq 100$ are both 3 mod 4 and 2 mod 5?

Exercise 4.3.4. Use the CRT to help determine all numbers $1 \leq n \leq 100$ such that $n \equiv 1 \bmod 5$ and $n \equiv 2 \bmod 7$.

Exercise 4.3.5. Let $m, n \geq 2$ be coprime. Show the restriction $\alpha : (\mathbb{Z}/mn\mathbb{Z})^\times \to (\mathbb{Z}/m\mathbb{Z})^\times \times (\mathbb{Z}/n\mathbb{Z})^\times$ satisfies $\alpha(1 + mn\mathbb{Z}) = (1 + m\mathbb{Z}, 1 + n\mathbb{Z})$ and α is multiplicative: $\alpha(ab + mn\mathbb{Z}) = \alpha(a + mn\mathbb{Z})\alpha(b + mn\mathbb{Z})$.

The next exercise may seem a bit contrived, but it can be viewed as an analogue of the highly useful Wilson's theorem to $n = pq$ and it is related to the trick we use for proving quadratic reciprocity.

Exercise 4.3.6. Let p, q be distinct odd primes. Let $P \in \mathbb{Z}/pq\mathbb{Z}$ be the product of all elements of $(\mathbb{Z}/pq\mathbb{Z})^\times$. Use the previous exercise together with Wilson's theorem to show $P \equiv 1 \bmod pq$. (*Hint:* Compute the product over $(\mathbb{Z}/p\mathbb{Z})^\times \times (\mathbb{Z}/q\mathbb{Z})^\times$ by first doing a product over $(\mathbb{Z}/p\mathbb{Z})^\times$, and then over $(\mathbb{Z}/q\mathbb{Z})^\times$.)

4.4 Quadratic Reciprocity

While the CRT provides nice closure to the problem of computing $\phi(n)$, our real goal is to apply it to quadratic reciprocity. It turns out that determining whether a is a square mod n essentially boils down to determining whether p is a square mod q, for primes p and q. Quadratic reciprocity says that, for odd primes p and q, whether p is a square mod q is determined by the reverse question of whether q is a square mod p. For the precise statement, the following notation will be helpful.

Definition 4.4.1. *Let p be an odd prime. The* **Legendre symbol**, *or* **quadratic residue symbol** $\pmod p$ *is defined for $a \in \mathbb{Z}$ by*

$$\left(\frac{a}{p}\right) = \begin{cases} 1 & a \text{ is a nonzero square } \bmod p \\ 0 & a \equiv 0 \bmod p \\ -1 & \text{else.} \end{cases}$$

So if a is relatively prime to p, then $\left(\frac{a}{p}\right)$ is 1 or -1, according to whether a is a square mod p or not. Note that $\left(\frac{a}{p}\right)$ depends only upon the congruence class of a mod p.

Example 4.4.1. For $p = 3$, we have $\left(\frac{0}{3}\right) = 0$, $\left(\frac{1}{3}\right) = 1$ and $\left(\frac{2}{3}\right) = -1$. (See Example 3.2.4.)

Example 4.4.2. For odd p, we have $\left(\frac{-1}{p}\right) = 1$ if $p \equiv 1 \bmod 4$ and $\left(\frac{-1}{p}\right) = -1$ if $p \equiv 3 \bmod 4$ by Lagrange's lemma (Lemma 4.1.4). Note we can write this in a uniform way as saying

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

In this formulation, Lagrange's lemma is also called the **first supplemental law to quadratic reciprocity**.

The following exercise says that for a prime to p, $\left(\frac{a}{p}\right)$ is 1 half of the time and -1 half of the time.

Exercise 4.4.1. Let p be an odd prime. Show that map $x \mapsto x^2$ on $(\mathbb{Z}/p\mathbb{Z})^\times$ is 2-to-1. Conclude that the number of squares in $(\mathbb{Z}/p\mathbb{Z})^\times$ is equal to the number of non-squares.

The usefulness of the Legendre symbol notation is because of the following result.

Proposition 4.4.2. *Let p be odd. The function $\left(\frac{\cdot}{p}\right)$ is (totally) multiplicative, i.e., for any $a, b \in \mathbb{Z}$, $\left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right)\left(\frac{b}{p}\right)$.*

Proof. Note that if $p \mid a$ or $p \mid b$, both sides of the equality are zero, so assume a, b are both coprime to p.

First suppose $\left(\frac{a}{p}\right) = 1$. Then $a \equiv x^2 \bmod p$ for some $x \in \mathbb{Z}$, $x \not\equiv 0 \bmod p$. It is easy to see that ab is a square mod p if and only if $ab(x^{-1})^2$ is a square mod p, but $ab(x^{-1})^2 \equiv b \bmod p$. Whence $\left(\frac{ab}{p}\right) = \left(\frac{b}{p}\right)$, which is the desired equality.

The same argument applies if $\left(\frac{b}{p}\right) = 1$, so we are reduced to treating the case that $\left(\frac{a}{p}\right) = \left(\frac{b}{p}\right) = -1$, where we need to show $\left(\frac{ab}{p}\right) = 1$. We can use a counting argument together with the previous exercise.

Assume $\left(\frac{a}{p}\right) = -1$. Note we can view multiplication by a as a map from $(\mathbb{Z}/p\mathbb{Z})^\times$ to itself: $x \mapsto ax$. Further, it is easy to see this is a bijection. By the case $\left(\frac{b}{p}\right) = 1$, we know ax is a non-square whenever x is a square. By the previous exercise, this must account for all $\frac{p-1}{2}$ times ax is a square as x ranges over $(\mathbb{Z}/p\mathbb{Z})^\times$. Thus if $x = b$ with b a non-square ($\left(\frac{b}{p}\right) = -1$), we have that $ax = ab$ must be a square, i.e., $\left(\frac{ab}{p}\right) = 1 = \left(\frac{a}{p}\right)\left(\frac{b}{p}\right)$. □

Consequently, if $q_1^{e_1} \cdots q_r^{e_r}$ is the prime-power factorization of $a \in \mathbb{N}$, to determine whether a is a square mod p (an odd prime), it suffices to determine whether each $\left(\frac{q_i}{p}\right)$ is 1 or -1 as

$$\left(\frac{a}{p}\right) = \left(\frac{q_1}{p}\right)^{e_1} \cdots \left(\frac{q_r}{p}\right)^{e_r}.$$

If a is even, one of these q_i's will be 2, but we can instead replace a with $a + p$ (or $a - p$ or $p - a$ or ...) which is odd, to assume each q_i is odd (or alternatively keep a the same

and compute as $\left(\frac{2}{p}\right) = \left(\frac{p+2}{p}\right)$, factoring $p+2$ into odd primes). So to determine whether a number is a square mod p, it suffices to determine $\left(\frac{q}{p}\right)$ for each odd prime q. Of course $\left(\frac{p}{p}\right) = 0$, so we may also assume $q \neq p$.

We will need one more auxiliary result to prove quadratic reciprocity, which in turn requires a basic fact about polynomials over fields.

Exercise 4.4.2. Let F be a field and $f(x)$ a polynomial of degree n over F, i.e., $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots a_1 x^1 + a_0$, where each $a_i \in F$ and $a_n \neq 0$.

(i) Prove that $x - b$ divides $f(x)$ (i.e., $f(x) = (x-b)g(x)$ for a polynomial $g(x)$ over F of degree $n-1$) if and only if $f(b) = 0$. (*Suggestion:* Use polynomial division and Fermat descent on the degree of $f(x)$.)

(ii) Conclude that there are at most n distinct roots of F.

Proposition 4.4.3. (Euler's criterion) *Let p be an odd prime and $a \in \mathbb{Z}$ be relatively prime to p. Then*

$$\left(\frac{a}{p}\right) \equiv a^{\frac{p-1}{2}} \bmod p.$$

Proof. Recall by Fermat's little theorem, we have $x^{p-1} \equiv 1 \bmod p$ for all x which are invertible mod p. So if a is a square mod p, i.e., $a \equiv x^2 \bmod p$ for some such x, then

$$\left(\frac{a}{p}\right) \equiv 1 \equiv x^{p-1} \equiv a^{\frac{p-1}{2}} \bmod p. \tag{4.4.1}$$

So it suffices to treat the case where a is not a square mod p.

Equivalently, we want to show if $a \in (\mathbb{Z}/p\mathbb{Z})^\times$ is not a square, then $a^{\frac{p-1}{2}} = -1$. (For the rest of the proof, we work in $\mathbb{Z}/p\mathbb{Z}$ rather than $\mathbb{Z}$.) Since $(a(p-1)/2)^2 = 1$, we always have $a^{(p-1)/2} = \pm 1$ since the only elements whose square is 1 in $\mathbb{Z}/p\mathbb{Z}$ are ± 1. (We've seen this in the proof of Wilson's theorem, as $x^2 = 1$ is equivalent to $x \in \mathbb{Z}/p\mathbb{Z}$ being its own inverse. Alternatively, apply the above exercise to the polynomial $f(x) = x^2 - 1$ over $F = \mathbb{Z}/p\mathbb{Z}$.) So it suffices to show $a^{(p-1)/2} \neq 1$ for any non-square $a \in (\mathbb{Z}/p\mathbb{Z})^\times$.

By the previous exercise, the polynomial $f(x) = x^{(p-1)/2} - 1$ over $F = \mathbb{Z}/p\mathbb{Z}$ has at most $\frac{p-1}{2}$ roots in $\mathbb{Z}/p\mathbb{Z}$. By (4.4.1) we know each square $a \in (\mathbb{Z}/p\mathbb{Z})^\times$ is a root of $f(x)$. But there are precisely $\frac{p-1}{2}$ squares in $(\mathbb{Z}/p\mathbb{Z})^\times$ by Exercise 4.4.1. Thus whenever a is a non-square, a is not a root of $f(x)$, i.e., $a^{(p-1)/2} \neq 1$. □

Exercise 4.4.3. Use Euler's criterion to give an alternative proof of Proposition 4.4.2.

As an aside, the ideas in the proof of Euler's criterion can also be used to determine the group structure of $(\mathbb{Z}/p\mathbb{Z})^\times$, something you would do in an algebra class. We won't use this in our course, but I'll leave it as:

Exercise 4.4.4. Prove that for any prime p, the group $(\mathbb{Z}/p\mathbb{Z})^\times$ is cyclic. (*Suggestion:* Try contradiction.)

Exercise 4.4.5. Use the previous exercise to show $(\mathbb{Z}/pq\mathbb{Z})^\times$ is cyclic for any distinct primes p, q.

Theorem 4.4.4. (Law of quadratic reciprocity) *Let p and q be distinct odd primes. Then*

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

In other words, $\left(\frac{q}{p}\right) = \left(\frac{p}{q}\right)$ unless $p \equiv q \equiv 3 \bmod 4$, in which case $\left(\frac{q}{p}\right) = -\left(\frac{p}{q}\right)$.

Sometimes, for symmetry, quadratic reciprocity is stated as $\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = (-1)^{\frac{p-1}{2}\cdot\frac{q-1}{2}}$. While this is a very practical result for computing $\left(\frac{a}{p}\right)$ (see below), the real beauty of it lies in the symmetry—it gives us a relation between squares mod p and squares mod q that seems completely miraculous. By this I mean, there is no obvious reason why p being a square mod q should affect whether q is a square mod p, but in fact one determines the other (once we know their congruence classes mod 4). Since there is no obvious reason why these are related, there is no simple direct proof—all known proofs either use some clever trickery or more advanced mathematics. Gauss devised several proofs to try to find a "good" reason why this law holds, and he was happiest with his third proof, which he viewed as the most simple, but it is still somewhat technical.[3] We'll give a different proof, which I think is easier to present, but it still involves some trickery. I first learned it from [Sti03], and it is a variant of Rousseau's proof (based on Gauss's fifth proof) published in 1991 [Rou91].

Proof. Let p, q be distinct odd primes. Set

$$S = \left\{1 \le x \le \frac{pq-1}{2} | \gcd(x, pq) = 1\right\},$$

so we may regard $(\mathbb{Z}/pq\mathbb{Z})^\times = S \cup -S$, where $-S = \{-x | x \in S\}$. We will consider $\prod_{x\in S} x$ both mod p and mod q.

Note that mod p, we can list the elements of S as $\frac{q-1}{2}$ full sequences $1, 2, \dots, p-1 \bmod p$ and the half sequence $1, 2, \dots, \frac{p-1}{2} \bmod p$, excluding the multiples $q, 2q, \dots, \frac{p-1}{2}q$ of q. E.g., if $p = 5$ and $q = 7$, then $S = \{1 \le x \le 17 : \gcd(x, 35) = 1\}$ which we can write in rows as

1	2	3	4	~~5~~
6	~~7~~	8	9	~~10~~
11	12	13	~~14~~	~~15~~
16	17			

corresponding to the 3 full sequences mod p and 1 half sequence mod p, where we've crossed out the numbers to be excluded.

[3] See Eisenstein's simplification of Gauss's third proof on Wikipedia: https://en.wikipedia.org/wiki/Proofs_of_quadratic_reciprocity

Hence

$$\prod_{x\in S} x \equiv ((p-1)!)^{\frac{q-1}{2}} \left(\frac{p-1}{2}\right)!/q^{\frac{p-1}{2}} \left(\frac{p-1}{2}\right)! \equiv (-1)^{\frac{q-1}{2}} \left(\frac{q}{p}\right) \bmod p,$$

where the second equivalence comes from Wilson's theorem, along with the fact that $1/q^{\frac{p-1}{2}} \equiv \pm 1 \equiv q^{\frac{p-1}{2}} \equiv \left(\frac{q}{p}\right) \bmod p$, and Euler's criterion. Similarly

$$\prod_{x\in S} x \equiv (-1)^{\frac{p-1}{2}} \left(\frac{p}{q}\right) \bmod q.$$

In other words, writing $\alpha(x) = (x \bmod p, x \bmod q)$ as the map $\alpha : (\mathbb{Z}/pq\mathbb{Z}) \to (\mathbb{Z}/p\mathbb{Z}) \times (\mathbb{Z}/q\mathbb{Z})$ from the statement of the Chinese Remainder theorem, we have

$$\prod_{x\in S} \alpha(x) \equiv \left((-1)^{\frac{q-1}{2}} \left(\frac{q}{p}\right), (-1)^{\frac{p-1}{2}} \left(\frac{p}{q}\right)\right) \bmod (p,q) \tag{4.4.2}$$

(Here we write $\bmod (p,q)$ to mean mod p in the first component and mod q in the second.)

Recall the Chinese Remainder Theorem says that α is a bijection of $(\mathbb{Z}/pq\mathbb{Z})^\times$ with $(\mathbb{Z}/p\mathbb{Z})^\times \times (\mathbb{Z}/q\mathbb{Z})^\times$. Since $(\mathbb{Z}/pq\mathbb{Z})^\times = S \cup -S$, this means that $\alpha(S) = \{\alpha(x) | x \in S\}$ contains exactly one of (a,b) and $(-a,-b)$ for each (a,b) in

$$T = \left\{(a,b) \in (\mathbb{Z}/p\mathbb{Z})^\times \times (\mathbb{Z}/q\mathbb{Z})^\times : 1 \le a \le p, 1 \le b \le \frac{q-1}{2}\right\},$$

and conversely for each $(a,b) \in \alpha(S)$ either in (a,b) or $(-a,-b)$ is in T. (Here we used that if $\alpha(x) = (a,b)$, then $\alpha(-x) = (-a,-b)$.) Hence

$$\begin{aligned}\prod_{x\in P} \alpha(x) \equiv \pm \prod_{(a,b)\in T} (a,b) &\equiv \pm \left((p-1)!^{\frac{q-1}{2}}, \left(\frac{q-1}{2}\right)!^{p-1}\right) \\ &\equiv \pm \left((-1)^{\frac{q-1}{2}}, \left(\frac{q-1}{2}\right)!^{p-1}\right) \bmod (p,q),\end{aligned}$$

where we used Wilson's theorem again in the last equivalence.

Note that

$$-1 \equiv (q-1)! \equiv 1 \cdot 2 \cdots \frac{q-1}{2} \cdot (-1)(-2)\cdots(-\frac{q-1}{2}) \equiv (-1)^{\frac{q-1}{2}} \left(\frac{q-1}{2}\right)!^2 \bmod q,$$

hence

$$\left(\frac{q-1}{2}\right)!^{p-1} \equiv \left(\left(\frac{q-1}{2}\right)!^2\right)^{\frac{p-1}{2}} \equiv \left((-1)(-1)^{\frac{q-1}{2}}\right)^{\frac{p-1}{2}} \equiv (-1)^{\frac{p-1}{2}} (-1)^{\frac{p-1}{2}\frac{q-1}{2}} \bmod q.$$

Thus

$$\prod_{x\in P} \alpha(x) \equiv \pm \left((-1)^{\frac{q-1}{2}}, (-1)^{\frac{p-1}{2}} (-1)^{\frac{p-1}{2}\frac{q-1}{2}}\right) \bmod (p,q). \tag{4.4.3}$$

Dividing (4.4.2) by (4.4.3), we get

$$(1,1) \equiv \pm \left(\left(\frac{q}{p}\right), (-1)^{\frac{p-1}{2}\frac{q-1}{2}} \left(\frac{p}{q}\right) \right) \mod (p,q).$$

Since p and q are odd, this means that both $\left(\frac{q}{p}\right)$ and $(-1)^{\frac{p-1}{2}\frac{q-1}{2}}\left(\frac{p}{q}\right)$ (which are both ± 1 in $\mathbb{Z}$) must both be $+1$ or both be -1, whence

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

which is precisely the Quadratic Reciprocity Law. □

One application is, if p is large, this lets us determine if something is a square mod p quite quickly, much faster than trying to compute all squares mod p.

Example 4.4.3. Determine if 15 is a square mod 103.

First, by multiplicativity

$$\left(\frac{15}{103}\right) = \left(\frac{3}{103}\right)\left(\frac{5}{103}\right).$$

Now by quadratic reciprocity, we have

$$\left(\frac{3}{103}\right) = -\left(\frac{103}{3}\right) = -\left(\frac{1}{3}\right) = -1$$

and

$$\left(\frac{5}{103}\right) = \left(\frac{103}{5}\right) = \left(\frac{3}{5}\right) = -1.$$

Thus $\left(\frac{15}{103}\right) = (-1)(-1) = 1$, so 15 is a square mod 103, even though we didn't determine what it's a square of.

Example 4.4.4. Determine if 94 is a square mod 101.

We could write $94 = 2 \cdot 47$ and try to compute $\left(\frac{2}{101}\right)$ and $\left(\frac{47}{101}\right)$. The latter we can use quadratic reciprocity for. There is a **second supplementary law** to compute $\left(\frac{2}{p}\right)$ as well, so this is possible, though we will not prove it in this course (see the next section for a statement). Instead we compute

$$\left(\frac{94}{101}\right) = \left(\frac{-7}{101}\right) = \left(\frac{-1}{101}\right)\left(\frac{7}{101}\right) = 1 \cdot \left(\frac{101}{7}\right) = \left(\frac{3}{7}\right) = -1,$$

using both the first supplementary law and quadratic reciprocity. Thus we see 94 is not a square mod 101.

Example 4.4.5. Determine for what primes p is 3 a square mod p.

We know 3 is a square mod 2 and mod 3, so it suffices to consider odd primes $p > 3$.

By quadratic reciprocity we have

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

Now $\left(\frac{p}{3}\right) = 1$ if $p \equiv 1 \bmod 3$ and is -1 if $p \equiv 2 \bmod 3$. On the other hand $(-1)^{(p-1)/2}$ is 1 if $p \equiv 1 \bmod 4$ and -1 if $p \equiv 3 \bmod 4$. So $\left(\frac{p}{3}\right) = 1$ if $p \equiv 1 \bmod 3$ and $p \equiv 1 \bmod 4$ or $p \equiv 2 \bmod 3$ and $p \equiv 3 \bmod 4$.

Hence 3 is a square mod p if and only if $p = 2, 3$ or $p \equiv 1, 11 \bmod 12$. (To combine a congruence mod 3 and and congruence mod 4 to one mod 12, you can either use the CRT or just check it by hand).

Exercise 4.4.6. Determine if 21 is a square mod 101. What about mod 103?

Exercise 4.4.7. Determine if 92 is a square mod 101. What about mod 103?

Exercise 4.4.8. Determine for what primes p we have 5 is a square mod p.

Exercise 4.4.9. Determine for what primes p we have 7 is a square mod p.

4.5 Numbers of the form $x^2 + dy^2$

Fermat not only studied what numbers are of the form x^2+y^2, but also considered questions like what numbers are of the form $x^2 + 2y^2$ and $x^2 + 3y^2$? (Geometrically, the case $x^2 + 2y^2$ corresponds to asking what numbers are the sums of 3 squares where at least 2 of the squares have the same size.) In this section, we'll take a brief look at the question: For fixed $d \in \mathbb{N}$, for which $n \in \mathbb{N}$ does

$$x^2 + dy^2 = n \tag{4.5.1}$$

have a solution for $x, y \in \mathbb{Z}$. (Geometrically, this is asking when is n a sum of $d+1$ squares where all or all but one of the squares have the same size.) This will show off some of the power of quadratic reciprocity, as well as give you a glimpse into a very beautiful and rich part of number theory that occupied many great minds since Fermat.

This is a special case of Gauss's theory of **binary quadratic forms**, which are polynomials of the form

$$Q(x, y) = ax^2 + bxy + cy^2$$

for some $a, b, c \in \mathbb{Z}$. (Here binary refers to the fact that we have two variables, and more generally a **quadratic form** is a polynomial which is a sum of terms that all have degree two, i.e., a homogeneous polynomial of degree 2.) In some sense, these are the simplest kinds of Diophantine equations in 2 variables beyond the linear ones $ax + by = n$. Here the

basic question is, given $Q(x,y)$ determine when $Q(x,y) = n$ has a solution, i.e., which n are **represented by** (or **of the form**) $Q(x,y)$? It turns out that a complete understanding of (4.5.1) involves looking at more general binary quadratic forms.

Without bringing in Gauss's general theory of binary quadratic forms, there are still many things we can say.[4] Here are a few simple general results.

Exercise 4.5.1. Show that n is represented by $x^2 + dy^2$, i.e., (4.5.1) has a solution if and only if n is a norm from $\mathbb{Z}[\sqrt{-d}]$, i.e., $n = N(\alpha)$ for some $\alpha \in \mathbb{Z}[\sqrt{-d}]$.

Corollary 4.5.1. (Composition law) *For $d > 0$, if m and n are represented by $x^2 + dy^2$, so is mn.*

Proof. Under the hypotheses, $m = N(\alpha)$ and $n = N(\beta)$ for some $\alpha, \beta \in \mathbb{Z}[\sqrt{-d}]$. Thus $mn = N(\alpha\beta)$ by multiplicativity of the norm. □

Consequently, just like for $x^2 + y^2 = n$, the most fundamental case of (4.5.1) should be when n is prime. The composition law doesn't exactly reduce the general problem to the case where n is prime. For instance if $n = pq$, we can say n is represented by $x^2 + dy^2$ if both p and q are, but it could happen that n is still represented by $x^2 + dy^2$ when p and q are not. As an example, $21 = 1^2 + 5 \cdot 2^2$, but neither 3 nor 7 are represented by $x^2 + 5y^2$. Recall, we used unique factorization of $\mathbb{Z}[i]$ to prove a converse to the composition law for $x^2 + y^2$ (if mn are sum of two squares and coprime, then m and n are each sums of two squares—Corollary 4.1.7), but this example shows the composition law doesn't have a converse for $x^2 + 5y^2$.

The failure of a converse to this composition law is related to the failure of unique factorization for $\mathbb{Z}[\sqrt{-5}]$, and can be explained by Gauss's theory of binary quadratic forms. In this case, both 3 and 7 are represented by $Q(x,y) = 2x^2 + 2xy + 3y^2$, and one can show that if coprime m and n are represented $Q(x,y)$, then mn is represented by $x^2 + 5y^2$, which gives a reason why 21 is of the form $x^2 + 5y^2$. Binary quadratic forms give a modified converse of the composition law, which says as a special case: if pq is represented by $x^2 + 5y^2$ then either both p and q are represented by $x^2 + 5y^2$ or both p and q are represented by $2x^2 + 2xy + 3y^2$. Thus to determine which numbers are of the form $x^2 + 5y^2$ we want to determine which primes are of the form $x^2 + 5y^2$ as well as which primes are of the form $2x^2 + 2xy + 3y^2$.

One can get some simple necessary conditions using modular arithmetic:

Exercise 4.5.2. Show that if p is represented by $x^2 + 5y^2$, then $p = 5$ or $p \equiv 1, 9 \bmod 20$. (Prove this directly—we give an alternative proof using quadratic reciprocity below.)

Exercise 4.5.3. Show that if p is represented by $2x^2 + 2xy + 3y^2$, then $p = 2$ or $p \equiv 3 \bmod 4$.

[4] You can look at my Number Theory II notest [Marb] and the references therein for more about binary quadratic forms.

In the former exercise, the necessary conditions turn out to be sufficient, but this requires much more work to prove. For the latter exercise, sufficient conditions turn out to be $p = 2$ or $p \equiv 3, 7 \bmod 20$.

In any case, for the rest of the section we will just focus on the problem: *which primes are of the form $x^2 + dy^2$?* Moreover, we will only focus on the much easier aspect of determining *necessary conditions*. The question about a complete characterization of primes of the form $x^2 + dy^2$ is studied in the beautiful (though somewhat advanced) book *Primes of the form $x^2 + ny^2$* by David Cox [Cox13].

Proposition 4.5.2. *Suppose a prime p is represented by $x^2 + dy^2$. Then the following must hold:*

(i) p is a square $\bmod d$; and

(ii) $-d$ is a square $\bmod p$.

Proof. (i) Reduce $p = x^2 + dy^2 \bmod d$ to get $p \equiv x^2 \bmod d$.

(ii) Suppose $p = x^2 + dy^2$ for some $x, y \in \mathbb{Z}$. Since p is not a square, we can take $0 < y < p$, and thus y and y^2 are invertible mod p. Now $x^2 + dy^2 \equiv 0 \bmod p$, means $x^2 \equiv -dy^2 \bmod p$, so

$$-d \equiv x^2 y^{-2} \equiv (xy^{-1})^2 \bmod p,$$

i.e., $-d$ is a square mod p, i.e., $\left(\frac{-d}{p}\right) = 1$. □

This proposition says that there are two congruence conditions for p to be of the form $x^2 + dy^2$. The first one is easy to directly check for a given d, the second one is less so.

Example 4.5.1. Say $d = 5$. The squares mod d are $0, 1, 4$. So if $p = x^2 + dy^2$, condition (i) from the above proposition says that we must have $p = 5$ or $p \equiv 1, 4 \bmod 5$.

Condition (ii) says that we also need -5 to be a square mod p. Calculating the squares mod p for primes up to 50 shows that condition (ii) is satisfied for $p = 2, 3, 5, 7, 23, 29, 41, 43, 47, \ldots$. It looks like, apart from $p = 5$, this these are the primes $p \equiv 3, 7 \bmod 20$. But how do we prove this?

The beauty of quadratic reciprocity is, *quadratic reciprocity lets us check a square condition mod p with a square condition mod d*: given d we can say p is not represented by $x^2 + dy^2$ if p lies in certain congruence classes mod m (here $m = d$ or $m = 4d$, as we will see in examples below). Assume $p \neq 2, 5$ so we can use the law of quadratic reciprocity. (Clearly $2 \neq x^2 + 5y^2$ for any $x, y \in Z$, and $5 = x^2 + y^2$ for $(x, y) = (0, \pm 1)$.)

Then condition (ii) is the statement that $\left(\frac{-5}{p}\right) = 1$. We compute $\left(\frac{-5}{p}\right) = \left(\frac{(-1)\cdot 5}{p}\right) = \left(\frac{-1}{p}\right)\left(\frac{5}{p}\right) = \left(\frac{-1}{p}\right)\left(\frac{p}{5}\right)$ by multiplicativity of the Legendre symbol and quadratic reciprocity with $q = 5$ (which is 1 mod 4). Recall Lagrange's lemma (i.e., the first supplementary law to quadratic reciprocity) says $\left(\frac{-1}{p}\right)$ is 1 if $p \equiv 1 \bmod 4$ and -1 if $p \equiv 3 \bmod 4$. Further, our calculation of squares mod 5 tells us that $\left(\frac{p}{5}\right)$ is 1 if $p \equiv 1, 4 \bmod 5$ and -1 if $p \equiv 2, 3 \bmod 5$.

Now $\left(\frac{-5}{p}\right) = 1$ if and only if $\left(\frac{-1}{p}\right)$ and $\left(\frac{p}{5}\right)$ are both $+1$ or are both -1. They are both $+1$ when $p \equiv 1 \bmod 4$ and $p \equiv 1, 4 \bmod 5$, i.e., $p \equiv 1, 9 \bmod 20$. They are both -1 when $p \equiv 3 \bmod 4$ and $p \equiv 2, 3 \bmod 5$, i.e., $p \equiv 3, 7 \bmod 20$.

Hence condition (ii) tells us that prime p is not of the form $x^2 + 5y^2$ unless $p = 2, 5$ or $p \equiv 1, 3, 7, 9 \bmod 20$. This proves part of Exercise 4.5.2, but doesn't rule out primes which are $3, 7 \bmod 20$. (Remark: the primes $p \equiv 3, 7 \bmod 20$ are represented by the related form $Q(x, y) = 2x^2 + 2xy + 3y^2$, which is related to why this approach does not rule them out.) However, condition (i) rules out $p \equiv 3, 7 \bmod 20$. This gives an "indirect" proof

of Exercise 4.5.2, which *derives* the congruence conditions mod 20, rather than telling you the congruence conditions seemingly out of nowhere (or guessing them from a lot of calculations).

In general, if $d = q_1^{e_1} \cdots q_r^{e_r}$ with q_i's distinct primes, we want to compute

$$\left(\frac{-d}{p}\right) = \left(\frac{-1}{p}\right)\left(\frac{q_1}{p}\right)^{e_1} \cdots \left(\frac{q_r}{p}\right)^{e_r}.$$

The first supplementary law tells us to compute $\left(\frac{-1}{p}\right)$ we look at p mod 4. If q_i is odd, we compute $\left(\frac{q_i}{p}\right)$ as $(-1)^{(p-1)(q_i-1)/4}\left(\frac{p}{q_i}\right)$ by quadratic reciprocity. If $q_i = 2$, we can use the following:

Proposition 4.5.3. (Second supplementary law to quadratic reciprocity *Let p be an odd prime. Then*

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

The proof is somewhat involved and we will not do it here, but we just gave the statement to give a more complete picture of the theory. In any case, one of the bottom lines is that to determine if p is of the form $x^2 + dy^2$, quadratic reciprocity and the supplementary laws tell us one should look at congruences mod $4d$ (in fact mod $4q_1 \cdots q_r$ suffices). Here we are using the Chinese Remainder Theorem to say we can rewrite a collection of congruence conditions mod 4, mod q_1, ..., mod q_r to congruence conditions mod $4q_1 \cdots q_r$. The factor of 4 here comes from needing to use the first (and sometimes second) supplementary law. (While the second supplementary law requires a congruence mod 8, it is only needed when d is even, so in the end mod $4d$ or $4q_1 \cdots q_r$ suffices.)

Here are some exercises and more remarks for primes of the form $x^2 + dy^2$ for a few small d.

Exercise 4.5.4. Use the supplementary laws and Proposition 4.5.2 to show that if $p = x^2 + 2y^2$ (has a solution over $\mathbb{Z}$), then $p = 2$ or $p \equiv 1, 3 \bmod 8$. (This is an "indirect" approach to Exercise 3.2.2.)

Exercise 4.5.5. Use quadratic reciprocity and Proposition 4.5.2 to show that if $p = x^2 + 3y^2$ (has a solution over $\mathbb{Z}$), then $p = 3$ or $p \equiv 1 \bmod 3$. (This is an "indirect" approach to Exercise 3.2.3.)

We note that Fermat showed the above conditions for p to be of the form $x^2 + 2y^2$ or $x^2 + 3y^2$ are in fact sufficient. One approach is to use the fact that $\mathbb{Z}[\sqrt{-2}]$ and $\mathbb{Z}[\zeta_3] \supset \mathbb{Z}[\sqrt{-3}]$ both have unique factorization.

Here is a special case where we easily get necessary and sufficient conditions for a prime to be of the form $x^2 + dy^2$, also known to Fermat.

Exercise 4.5.6. (i) Show that if $p = x^2 + y^2$ and $p \neq 2$, then one of x, y must be even.
(ii) Use Fermat's two square theorem to prove that p is of the form $x^2 + 4y^2$ if and only if $p \equiv 1 \bmod 4$.

You might think that use of quadratic reciprocity is actually making things more complicated than what we did in Chapter 3, but that is only because (i) we've just explored things for small d so far where things are especially simple, and (ii) in Chapter 3 I already told you for what m you should look at $x^2 + dy^2 \bmod m$. As explained above, it's really quadratic reciprocity that tells us in general for what m we want to look at congruence conditions for numbers of the form $x^2 + dy^2$.

Exercise 4.5.7. Use quadratic reciprocity (and both supplementary laws) and Proposition 4.5.2 to determine congruence conditions for when $p = x^2 + 6y^2$ can have a solution.

Exercise 4.5.8. Use quadratic reciprocity (and both supplementary laws) and Proposition 4.5.2 to show that $p = x^2+14y^2$ can only have a solution if $p \equiv 1, 9, 15, 23, 25, 39 \bmod 56$.

What is interesting about the last exercise is that this is one of the first examples where the necessary congruence condition on p is *not* sufficient to guarantee p is of the form $x^2 + 14y^2$. The general theory say that one also needs to check a condition mod p. In this particular case, p is of the form $x^2 + 14y^2$ if and only if $p \equiv 1, 9, 15, 23, 25, 39 \bmod 56$ and the equation $(a^2 + 1)^2 \equiv 8$ has a solution (in a) mod p. On the other hand, what is true (via Gauss's theory of binary quadratic forms) is that p is of the form $x^2 + 14y^2$ or of the form $2x^2 + 7y^2$ if and only if $p \equiv 1, 9, 15, 23, 25, 39 \bmod 56$.

4.6 Sums of three and four squares

Another way to generalize Fermat's two square problem is to ask what numbers are sums of k squares for $k > 2$. As mentioned in the introduction, the answer is all positive integers are sums of k squares when $k \geq 4$ by:

Theorem 4.6.1. (Lagrange's four square theorem, 1770) *Every natural number is a sum of four squares, i.e., $n = x^2 + y^2 + z^2 + w^2$ has a solution with $x, y, z, w \in \mathbb{Z}$ for all $n \in \mathbb{N}$.*

The case of two squares is harder than that of two square or four squares, but of course the great Gauss could solve it in his *Disquisitiones* when he was 21:

Theorem 4.6.2. (Gauss's three square theorem, 1801, aka Legendre's three square theorem)[5] *Any $n \in \mathbb{N}$ is a sum of three squares, i.e., $n = x^2 + y^2 + z^2$ has a solution with $x, y, z \in \mathbb{Z}$, if and only if n is not of the form $4^k(8m + 7)$.*

[5] Some people, including me in the past, attribute this to Legendre. He certainly claimed he had a proof, though my current understanding is his proof was not correct. At least Gauss asserted there were serious issues with his proof.

We won't prove Gauss's three square theorem (Gauss used binary quadratic forms) but will indicate how to prove part of it. Here is the easy direction:

Proposition 4.6.3. *If n is a sum of three squares, then n is not of the form $4^k(8m+7)$.*

Proof. Suppose $n = x^2+y^2+z^2$ for some $x, y, z \in \mathbb{Z}$ but $n = 4^k(8m+7)$. Note if $k = 0$, we already know $8m+7$ is not a sum of three squares by Exercise 3.2.4.

So we must have $k \geq 1$. Since the squares mod 4 are just 0 and 1, for $x^2+y^2+z^2 \equiv 0 \bmod 4$ we need x, y, z all even. (This is similar to part of Proposition 3.2.5.) Then $\frac{n}{4} = 4^{k-1}(8m+7) = (\frac{x}{2})^2 + (\frac{y}{2})^2 + (\frac{z}{2})^2$ is also sum of three squares. By descent on k, we conclude that $8m+7$ is a sum of three squares, which is a contradiction by Exercise 3.2.4. □

For the hard direction, we can at least explain how it follows when $n = p$ is prime. Suppose p is not of the form $4^k(8m+7)$. Since $4 \nmid p$, this just means $p \not\equiv 7 \bmod 8$. If $p = 2$, this is obvious. Otherwise we have $p \equiv 1, 3, 5 \bmod 8$. If $p \equiv 1, 5 \bmod 8$, then $p \equiv 1 \bmod 4$, so $p = x^2+y^2$ for some x, y, hence $p = x^2+y^2+z^2$ with $z = 0$. If $p \equiv 3 \bmod 8$ (or $1 \bmod 8$), then a result of Fermat we mentioned after Exercise 4.5.4 but did not prove says $p = x^2+2y^2$ for some x, y. Then $p = x^2+y^2+z^2$ with $z = y$. This yields Gauss's three square theorem in the case n is prime.

Then one might hope to use a composition law and some kind of converse, as in the case of sums of two squares, to get the general case. However, it is easy to see that this is not possible: if we have two primes $p \equiv 3 \bmod 8$ and $q \equiv 5 \bmod 8$, then they are both sums of three squares by the last paragraph, but their product $pq \equiv 15 \equiv 7 \bmod 8$, so is not a sum of three squares by the above proposition/Exercise 3.2.4. Hence there is no composition law in general.

However, there is a composition law for sums of four squares, which helps makes proving Lagrange's four square theorem much easier than Gauss's three square theorem. We will explain this now.

For sums of two squares, recall we proved the composition law by using the norm map on $\mathbb{Z}[i]$, and the fact that this is multiplicative. The norm map on $\mathbb{Z}[i]$ (or any imaginary quadratic ring) is simply the restriction of the (algebraic) norm map $z \mapsto z\overline{z} = |z|^2$ from $\mathbb{C}$ to $\mathbb{R}$ to our quadratic ring. (Here $\overline{z}$ denotes the complex conjugate of z.) Recall also that multiplication by $z = re^{i\theta}$ ($r \geq 0$, $\theta \in \mathbb{R}$) in $\mathbb{C}$ acts geometrically on the complex plane by radial scaling by r and rotation about 0 by θ radians.

In the first half of the 19th century, William Rowan Hamilton tried to come up with an algebraic structure (e.g., a field) analogous to $\mathbb{C}$ which is 3-dimensional over $\mathbb{R}$, in the hopes that one could understand rotations in $\mathbb{R}^3$ algebraically. Eventually, he realized that this is not possible (the fact that it is impossible is related to the fact that there is no composition law for sums of three squares), but it is possible in 4 dimensions! However, one has to work with an algebraic structure which is non-commutative.

Definition 4.6.4. *We define* **Hamilton's quaternions** $\mathbb{H}$ *to be the four-dimensional vector space with basis* $\{1, i, j, k\}$,

$$\mathbb{H} = \{x + yi + zj + wk : x, y, z, w \in \mathbb{R}\},$$

together with an associative (vector) multiplication law which is $\mathbb{R}$-linear and satisfies

$$i^2 = j^2 = k^2 = ijk = -1.$$

We now explain what we mean by the multiplication law. First consider multiplication of basis elements. Multiplication of anything by 1 (the basis element 1 is the same as the real number 1) should be itself: $1 \cdot \alpha = \alpha \cdot 1 = \alpha$. The rules $i^2 = j^2 = k^2 = -1$ are evident: i, j and k are (distinct) square roots of $-1 \in \mathbb{R}$. (Hence we can think of $\mathbb{R} \oplus \mathbb{R}i$, $\mathbb{R} \oplus \mathbb{R}j$ and $\mathbb{R} \oplus \mathbb{R}k$ as distinct subspaces which are all algebraically the same as $\mathbb{C}$.) These rules combined with $ijk = -1$ then tells us how to multiply any two basis elements—e.g.

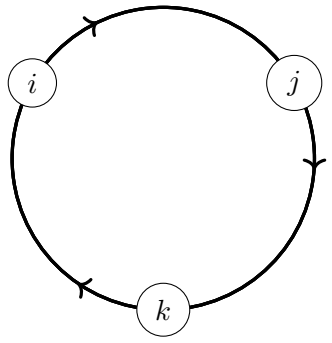
$$(ijk)k = (-1)k \implies ij(k^2) = -ij = -k \implies ij = k.$$

(Here we used that -1 commutes with each basis element, which is part of what I mean by multiplication being "$\mathbb{R}$-linear".) The following exercise tells us most of the other cases of multiplication of basis elements (with the rest being similar, which we explain with a picture below).

Exercise 4.6.1. Show $jk = i$, $ki = j$ and $ji = -ij = -k$.

In particular, the order of multiplication matters: $ij \neq ji$!

To make things easier to remember, we can visualize the multiplication table for i, j, k with the following picture:



The way to interpret this is as follows. Any of i, j, k square is -1, so say we want to multiply two distinct elements of $\{i, j, k\}$. The product will always be plus or minus the other element of $\{i, j, k\}$, and if the order of multiplication agrees with the direction of arrows in the picture, the sign is $+$, but if it disagrees, then the sign is $-$. For instance, when we multiply j and k, we will get $\pm i$. If we multiply them in the order jk, we get $+i = i$, and if we multiply them in the order kj we get $-i$.

Now we can extend this multiplication of basis elements $\{1, i, j, k\}$ to multiplication of elements of $\mathbb{H}$ in a way that is $\mathbb{R}$-linear: if $\alpha = x + yi + zj + wk$ and $\alpha' = x' + y'i + z'j + w'k$, we consider their product to be

$$\begin{aligned}\alpha\alpha' = &(x + yi + zj + wk)(x' + y'i + z'j + w'k)\\ = \;& xx' + xy'i + xz'j + x'w'k\\ &+ yx'i + yy'i^2 + yz'ij + yw'ik\\ &+ zx'j + zy'ji + zz'j^2 + zw'jk\\ &+ wx'k + wy'ki + wz'kj + ww'k^2.\end{aligned}$$

That is, we just distribute, and we are allow to commute the real numbers x, y, z, w and x', y', z', w', but we are not allowed to commute two (distinct) basis elements $\{i, j, k\}$. Then we just compute the products of the basis elements, and we can then rewrite

$$\alpha\alpha' = x'' + y''i + z''j + w''k,$$

for some $x'', y'', z'', w'' \in \mathbb{R}$. For instance, the x'' term will come from the product of basis elements of the form 1^2, i^2, j^2 and k^2, giving

$$x'' = xx' - yy' - zz' - ww'.$$

Exercise 4.6.2. In the notation above, determine y'' in terms of x, y, z, w and x', y', z', w'.

Example 4.6.1. Let $\alpha = 1 + 2i + 3j$, $\beta = 4 + 5i + 7k$. Then

$$\begin{aligned}\alpha\beta &= 1(4 + 5i + 7k) + 2i(4 + 5i + 7k) + 3j(4 + 5i + 7k)\\ &= (4 + 5i + 7k) + (8i + 10i^2 + 14ik) + (12j + 15ji + 21jk)\\ &= (4 + 5i + 7k) + (8i - 10 - 14j) + (12j - 15k + 21i)\\ &= (4 - 10) + (5 + 8 + 21)i + (-14 + 12)j + (7 - 15)k = -6 + 34i - 2j - 8k.\end{aligned}$$

Exercise 4.6.3. In the above example of $\alpha = 1 + 2i + 3j$, $\beta = 4 + 5i + 7k$, compute the product in the reverse order: $\beta\alpha$.

Now we can add and multiply any two elements of $\mathbb{H}$. (Contrast this with arbitrary real vector spaces, where you can only add vectors and multiply a vector with a scalar.) It is not too hard to check the following:

Proposition 4.6.5. *$\mathbb{H}$ is what is known as a* **skew field** *or a* **division ring**, *i.e., it satisfies all 6 field axioms with the sole exception of commutativity of multiplication.*

We remark that the terms skew field and division ring are interchangeable, with division ring probably being more widely used now. However, I think the term skew field is maybe more helpful to use when you are first seeing $\mathbb{H}$ to emphasize that it is like a field, i.e., like $\mathbb{R}$ or $\mathbb{C}$, only not commutative. We also remark that a structure R satisfying the 5 ring axioms with the possible exception of commutativity of multiplication is called a **noncommutative ring**. Skew fields (i.e., division rings) are special cases of noncommutative rings. When $n \geq 2$, the set of $n \times n$ matrices $M_n(\mathbb{R})$ (or $M_n(\mathbb{C})$) is an example of a noncommutative ring which is not a division ring.

Exercise 4.6.4. Show $M_2(\mathbb{R})$ is not a division ring. (Thus $M_2(\mathbb{R})$ is a different 4-dimensional algebraic structure than $\mathbb{H}$.)

Definition 4.6.6. *For $\alpha = x + yi + zj + wk \in \mathbb{H}$, we define the* **conjugate** *of α to be*

$$\overline{\alpha} = x - yi - zj - wk,$$

and the **norm** *of α to be*

$$N(\alpha) = \alpha\overline{\alpha} = x^2 + y^2 + z^2 + w^2.$$

Thus the norm map is defined $N : \mathbb{H} \to \mathbb{R}_{\geq 0}$.

Note if we think of $\mathbb{C}$ as any of the following subsets of $\mathbb{H}$: $\{x + yi : x, y \in \mathbb{R}\}$, $\{x + yj : x, y \in \mathbb{R}\}$ or $\{x + yk : x, y \in \mathbb{R}\}$, then the norm on $\mathbb{H}$ matches the norm on $\mathbb{C}$, e.g., $N(x+yi) = x^2+y^2$.[6]

Exercise 4.6.5. Check that for $\alpha = x + yi + zj + wk \in \mathbb{H}$, we indeed have $\alpha\overline{\alpha} = x^2 + y^2 + z^2 + w^2$.

Exercise 4.6.6. Let $\alpha = x + yi + zj + wk$ and $\beta = x' + y'i + z'j + w'k$ in $\mathbb{H}$, and write $\alpha = x + \alpha_0$, $\beta = x' + \beta_0$ where $\alpha_0 = yi + zj + wk$ and $\beta_0 = y'i + z'j + w'k$.
(i) Show $\overline{\alpha_0\beta_0} = \overline{\beta_0} \cdot \overline{\alpha_0}$.
(ii) Deduce that $\overline{\alpha\beta} = \overline{\beta} \cdot \overline{\alpha}$.

Proposition 4.6.7. *Let $\mathbb{Z}[i, j, k] = \{x + yi + zi + wk \in \mathbb{H} : x, y, z, w \in \mathbb{Z}\}$.*
(i) For $n \in \mathbb{Z}$, we have n is a sum of four squares if and only if n is a norm from $\mathbb{Z}[i, j, k]$.
(ii) **(Composition law)** *If m and n are sums of four squares, so is mn.*

Proof. (i) This is obvious as $N(x + yi + zi + wk) = x^2 + y^2 + z^2 + w^2$.
(ii) Suppose m and n are sums of four squares, so $m = N(\alpha)$ and $n = N(\beta)$ for some $\alpha, \beta \in \mathbb{Z}[i, j, k]$. Then from the previous exercise

$$N(\alpha\beta) = \alpha\beta\overline{\alpha\beta} = \alpha(\beta\overline{\beta})\overline{\alpha} = N(\beta)\alpha\overline{\alpha} = N(\alpha)N(\beta).$$

(Here we used the fact that $N(\alpha), N(\beta) \in \mathbb{R}$, so they commute with everything.) Thus the norm map is multiplicative, which implies (ii) in light of (i). □

Now by the composition, proving Lagrange's four square theorem reduces to the following:

Proposition 4.6.8. *Let $p \in \mathbb{N}$ be prime. Then p is a sum of four squares.*

[6] There are other square roots of -1 in $\mathbb{H}$, infinitely many in fact, by taking appropriate combinations of $1, i, j, k$. (*Hint:* To prove this, start looking among elements of norm 1.) Thus there are infinitely many ways to realize $\mathbb{C}$ as a subset of $\mathbb{H}$. In any of these realizations, conjugation (and thus norm) on $\mathbb{H}$ agrees with conjugation (and thus norm) on $\mathbb{C}$.

The proof will use the following fact, which we will take for granted. As in the case of commutative rings, we will call an nonzero element of $u \in \mathbb{Z}[i,j,k]$ a **unit** if the inverse of $u \in \mathbb{H}$ also lies in $\mathbb{Z}[i,j,k]$. It is easy to see from multiplicativity of the norm on $\mathbb{H}$ that u being a unit is equivalent to $N(u) = 1$. We say a non-zero nonunit $\alpha \in \mathbb{Z}[i,j,k]$ is **reducible** if there exist $\beta, \gamma \in \mathbb{Z}[i,j,k]$ which are both nonzero non-units such that $\alpha = \beta\gamma$, and **irreducible** otherwise.

Theorem 4.6.9. *The non-commutative ring $\mathbb{Z}[i,j,k]$ satisfies the following weak prime divisor property: if $\pi \in \mathbb{Z}[i,jk]$ is an irreducible with odd norm, and $\pi \mid \alpha\beta$ with $\alpha, \beta \in \mathbb{Z}[i,j,k]$, then $\pi \mid \alpha$ or $\pi \mid \beta$.*

(I haven't exactly said what I mean by $\pi \mid \alpha$—some thought is merited since $\mathbb{Z}[i,j,k]$ is non-commutative—e.g., left divisors versus right divisors. But we'll only apply this notion to $p \mid \alpha$ with $p \in \mathbb{N}$ below, and since p commutes with everything in $\mathbb{Z}[i,j,k]$, this is not an issue.)

The condition that π has odd norm really is necessary. One can see this from looking at the example $\pi = 1 + i$, $\alpha = 1 + j$ and $\beta = \overline{\alpha} = 1 - j$. These are all irreducible in $\mathbb{Z}[i,j,k]$. Then $\alpha\beta = N(\alpha) = 1^2 + 1^2 = 2$, and since also $2 = \pi\overline{\pi}$, we have $\pi \mid 2 = \alpha\beta$. But one can show that $\pi \nmid \alpha$ and $\pi \nmid \beta$.

Proof of Proposition. Clearly $2 = 1^2 + 1^2 + 0^2 + 0^2$, so assume p is odd. Then, because squaring is a 2-to-1 map on $(\mathbb{Z}/p\mathbb{Z})^\times$, including 0 mod p there are precisely $\frac{p+1}{2}$ squares mod p (this was Exercise 4.4.1). Hence the map $x \mapsto -1 - x^2$ on $\mathbb{Z}/p\mathbb{Z}$ takes exactly $\frac{p+1}{2}$ values. But as there only $\frac{p-1}{2}$ non-squares mod p, at least one of these values must be a square. That is, for some $x, y \in \mathbb{Z}$, $-1 - x^2 \equiv y^2 \bmod p$, i.e., $p \mid (x^2 + y^2 + 1)$. (This fact is another lemma of Lagrange, similar to Lemma 4.1.4.)

For x, y as above, consider $\alpha = x + yi + j \in \mathbb{Z}[i,j,k]$. Then

$$p \mid (x^2 + y^2 + 1) = N(\alpha) = \alpha\overline{\alpha} = (x + yi + j)(x - yi - j).$$

Note that $\frac{x \pm yi \pm j}{p} \notin \mathbb{Z}[i,j,k]$, hence p does not divide α or $\overline{\alpha}$. By the (weak) prime divisor property in $\mathbb{Z}[i,j,k]$, this means that p must be reducible in $\mathbb{Z}[i,j,k]$, so we can write $p = \beta\gamma$ for β, γ nonzero non-units. Then

$$p^2 = N(p) = N(\beta\gamma) = N(\beta)N(\gamma).$$

Since $N(\beta)$ and $N(\gamma)$ are positive integers greater than 1, and p is prime in $\mathbb{N}$, we must have $N(\beta) = N(\gamma) = p$. In particular, p is a norm from $\mathbb{Z}[i,j,k]$, so a sum of four squares by the previous proposition. □

The numbers in $\mathbb{Z}[i,j,k]$ are called the **Lipschitz integers**. There is actually a larger set of "quaternion integers" one can work with, the **Hurwitz integers**

$$\left\{ \frac{x + yi + zj + wk}{2} : x, y, z, w \in \mathbb{Z},\ x \equiv y \equiv z \equiv w \bmod 2 \right\}.$$

(The Hurwitz integers are obtained from the Lipschitz integers by adjoining the element $\frac{1+i+j+k}{2}$.) One way of proving the Lipschitz integers satisfy the above weak prime divisor

property is to first prove the Hurwitz integers satisfy the usual prime divisor property (i.e., one need not assume $N(\pi)$ is odd). This can be done by showing the Hurwitz integers possess the division property (and thus a Euclidean algorithm). (The Lipschitz integers do not satisfy the division property.) Consequently, sometimes people say that the Hurwitz integers have "unique factorization," but one needs to be more careful what one means because the order of factorization matters. See, e.g., [CS03]. (We could also present the above proof in terms of Hurwitz integers rather than Lipschitz integers, as in [Sti03] or [Mara, Ch 8].)

Anyway, we will not prove Theorem 4.6.9. The point of the above was to show that one can prove Lagrange's four square theorem using a similar approach to our proof of Fermat's two square theorem. Also, the quaternions are an interesting mathematical object. It turns out they can be used to achieve Hamilton's original goal of getting an algebraic way of treating 3-dimensional geometry. In particular, they provide an algebraic way of studying rotations in $\mathbb{R}^3$, and thus are useful in physics and engineering. One can also use the quaternions to give a proof of Gauss's three square theorem.

There are various other proofs of the four square theorem. For instance, in my earlier notes [Mara, Ch 8] I sketch out a geometric proof as well as an analytic proof.

You might also wonder about other algebraic structures beyond quaternions giving more composition laws. In fact there are such structures. For instance, there are the 8-dimensional **octonions**, which allow one to prove a composition law for sums of 8 squares. However, similar to how we lost commutativity going from $\mathbb{C}$ to $\mathbb{H}$, when you move to the octonions, you lose associativity (which is bad, but not quite as bad as it sounds at first).

Exercise 4.6.7. In the above proof we only worked with α of the form $x + yi + zj$, which has norm $x^2 + y^2 + z^2$. Why does the argument not imply that any prime p is a sum of three squares?

Exercise 4.6.8. How many units are there in the Lipschitz integers? What about the Hurwitz integers?

Chapter 5

Pell's Equation

One of the earliest issues grappled with in number theory is the fact that geometric quantities are often not rational. For instance, if we take a right triangle with two side lengths equal to 1, the hypopotamus has length $\sqrt{2}$, which is irrational. But how can we do arithmetic with irrational numbers? Well, perhaps the most basic thing is to work with rational approximations. Almost 4000 years ago, Babylonians had discovered the following approximation to $\sqrt{2}$:

$$\sqrt{2} = 1.41421356... \approx \frac{30547}{21600} = 1.41421\overline{296}. \tag{5.0.1}$$

In this chapter we'll explain how find the (integer) solutions to **Pell's equation**:

$$x^2 - dy^2 = 1, \tag{5.0.2}$$

and how gives us good approximations to $\sqrt{d}$ (see Proposition 5.2.3).

Following Stigler's law of eponomy[1], Pell's equation was studied by the Indian mathematician and astronomer Brahmagupta in 628 (who discovered the composition law Proposition 5.1.1) and with a general method of solution by another Indian mathematician and astronomer, Bhaskara II, in 1150. In Europe, methods for solving Pell's equation were rediscovered hundreds of years later by Fermat and Lord Brouncker. Euler misattributed Lord Brouncker's solution after reading a discussion of Lord Brouncker's method written by the English mathematician John Pell (1611–1685).

Throughout this chapter $d > 1$ is a positive integer which is not a square.

5.1 Units and Pell's equation

Recall that a unit u of $\mathbb{Z}[\sqrt{d}]$ was defined to be an element such that u has a multiplicative inverse $u^{-1} \in \mathbb{Z}[\sqrt{d}]$ (i.e., the real number $u \neq 0$ and $\frac{1}{u} \in \mathbb{Z}[\sqrt{d}]$). Further, by Lemma 2.1.2, we know that $x + y\sqrt{d} \in \mathbb{Z}[\sqrt{d}]$ $(x, y, \in \mathbb{Z})$ is a unit if and only if

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

[1]That no scientific discovery is named after it's first discoverer. The Pythagorean theorem is another famous example. Of course there are many counterexamples to Stigler's law as well. Appropriately, Stigler's law itself is not.

Thus solutions to Pell's equation (5.0.2) are in natural bijection with the units of $\mathbb{Z}[\sqrt{d}]$ with norm 1.

On the other hand, we also know by Proposition 3.3.4 that the set of units

$$U = U_d = \mathbb{Z}[\sqrt{d}]^\times$$

of $\mathbb{Z}[\sqrt{d}]$ form an (abelian) group. We also denote by $U^+ = U_d^+$ the set of units in $U = U_d$ of norm 1, so we can think of the solutions to Pell's equation as the subgroup U^+ of U.

Exercise 5.1.1. Check that U^+ is indeed a subgroup of U.

This group structure will help us determine the set of solutions to Pell's equation. First, we have the following, which is similar to the composition law for sums of two squares.

Proposition 5.1.1. (Composition law) *If (x_1, y_1) and (x_2, y_2) are solutions to*

$$x_1^2 - dy_1^2 = m, \quad x_2^2 - dy^2 = n.$$

Then the composition of these solutions defined by

$$(x_3, y_3) = (x_1, y_1) \cdot (x_2, y_2) := (x_1x_2 + dy_1y_2, x_1y_2 + y_1x_2) \tag{5.1.1}$$

is a solution of

$$x_3^2 - dy_3^2 = mn.$$

Proof. We simply translate the above into a statement about norms. The hypothesis says $N(x_1 + y_1\sqrt{d}) = m$ and $N(x_2 + y_2\sqrt{d}) = n$. Now observe that

$$(x_1 + y_1\sqrt{d})(x_2 + y_2\sqrt{d}) = x_1x_2 + ny_1y_2 + (x_1y_2 + y_1x_2)\sqrt{d} = x_3 + y_3\sqrt{d}.$$

Hence by the multiplicative property of the norm,

$$x_3^2 - dy_3^2 = N(x_3 + y_3\sqrt{d}) = N(x_1 + y_1\sqrt{d})N(x_2 + y_2\sqrt{d}) = mn.$$

□

In particular, when $m = n = 1$, this says that we can compose two solutions to Pell's equation to get a third solutions. We can also compose two solutions to $x^2 - dy^2 = -1$ to get a solution to $x^2 - dy^2 = +1$. Both of these are summarized in this corollary.

Corollary 5.1.2. *Let $u_1 = x_1 + y_1\sqrt{d}$ and $u_2 = x_2 + y_2\sqrt{d}$ be units of $\mathbb{Z}[\sqrt{d}]$. If $N(u_1) = N(u_2)$, then the composition $(x_1, y_1) \cdot (x_2, y_2)$ defined in Eq. (5.1.1) also a solution to Pell's equation (5.0.2).*

The following should be obvious if you've had an algebra class, but since we never covered isomorphisms, I'm not entirely sure if this is obvious to you:

Exercise 5.1.2. Let $G \subset \mathbb{Z} \times \mathbb{Z}$ be the set of solutions to Pell's equation (5.0.2). Show that the composition law Eq. (5.1.1) makes G into a group. (Note the above corollary just says composition is a binary operation on G.)

Now that we know the composition law, the hope is that if we can determine a few good solutions to Pell's equation, then maybe we can generate all solutions by composing those we know. Moreover, ideally these good solutions should be the "smallest nontrivial" solutions to Pell's equation. By the **trivial solutions** to Pell's equation, we mean the obvious ones: $(x, y) = (\pm 1, 0)$, which correspond to the elements of U^+ which lie in $\mathbb{Z}$, i.e., ± 1.

Example 5.1.1. Consider $d = 2$. Note $(1, 1)$ is a solution to $x^2 - 2y^2 = -1$. The composition $(1, 1) \cdot (1, 1) = (3, 2)$ is a nontrivial solution to Pell's equation $x^2 - 2y^2 = 1$. Similarly, we compute $(3, 2) \cdot (3, 2) = (17, 12)$ and $(3, 2) \cdot (17, 12) = (99, 70)$.

Hence $(3, 2)$, $(17, 12)$ and $(99, 70)$ are three nontrivial solutions to $x^2 - 2y^2 = 1$.

Exercise 5.1.3. Find a nontrivial solution to $x^2 - 3y^2 = 1$. Use composition to find two more (distinct) solutions to $x^2 - 3y^2 = 1$.

We remark that in the case of $d = 3$, unlike $d = 2$, there are no units of norm -1. In fact, the following more general statement is true.

Exercise 5.1.4. Suppose $d \equiv 3 \bmod 4$. Show $\mathbb{Z}[\sqrt{d}]$ has no units of norm -1, i.e., $U_d = U_d^+$.

The converse to the previous exercise does not hold, i.e., there may or may not be a unit of norm -1 when $d \not\equiv 3 \bmod 4$. We've seen there is such a unit when $d = 2$. The next exercise gives you an example where there isn't a unit of norm -1 but $d \not\equiv 3 \bmod 4$.

Exercise 5.1.5. Show that $\mathbb{Z}[\sqrt{6}]$ has no units of norm -1.

In general, it is an open problem to determine for what d there are units of norm -1 in $\mathbb{Z}[\sqrt{d}]$. It's not clear that there is a nice answer to this problem, but there results about how often $\mathbb{Z}[\sqrt{d}]$ has units of norm -1.

5.2 Approximation and existence of solutions

At the end of the last section we saw that there are nontrivial solutions to Pell's equation when $d = 2, 3$. Next we will prove the existence of a non-trivial solution for all non-square d, which is originally due to Lagrange in 1768. However, the proof we will give is due to Dirichlet (ca. 1840). It uses the pigeonhole principle. You've probably at least seen the finite version in your Discrete Math class.

Pigeonhole principle

- (finite version) If $m > k$ pigeons go into k boxes, at least one must box must contain more than 1 pigeon.
- (infinite version) If infinitely many pigeons go into k boxes, at least one box must contain infinitely many pigeons).

Proposition 5.2.1. (Dirichlet's approximation theorem) *For any non-square $d > 1$ and integer $B > 1$, there exist $a, b \in \mathbb{N}$ such that $b < B$ and*

$$|a - b\sqrt{d}| < \frac{1}{B}.$$

This says that $|\frac{a}{b} - \sqrt{d}| < \frac{1}{bB} \frac{1}{B}$, which is a precise way of saying $\frac{a}{b}$ is close to $\sqrt{d}$. E.g., it says we can find a rational approximation $\frac{a}{b}$ for $\sqrt{2}$ which is accurate within $\frac{1}{100,000}$ (so $\frac{a}{b}$ and $\sqrt{2}$ agree to 5 decimal places, after rounding if necessary[2]) with denominator $b < 100,000$. Such an example was exhibited at the beginning of this chapter in (5.0.1).

Proof. Consider the $B - 1$ irrational numbers

$$\sqrt{d}, 2\sqrt{d}, \dots, (B-1)\sqrt{d}.$$

For each such number $k\sqrt{d}$ $(1 \leq k \leq B-1)$, let $a_k \in \mathbb{N}$ be such that

$$0 < a_k - k\sqrt{d} < 1.$$

Partition the interval $[0, 1]$ into B subintervals of length $\frac{1}{B}$. Then, of the $B + 1$ numbers

$$0, a_1 - \sqrt{d}, a_2 - \sqrt{d}, \dots, a_{B-1} - (B-1)\sqrt{d}, 1$$

in $[0, 1]$ two of them must be in the same subinterval of length $\frac{1}{B}$. Hence they are less than distance $\frac{1}{B}$ apart, i.e., their difference satisfies $|a - b\sqrt{d}| < \frac{1}{B}$. Further their irrational parts must be distinct, so we have $-B < b < B$ with $b \neq 0$. If $b > 0$ we are done; if $b < 0$, simply multiply a and b by -1. Clearly we need $a > 0$ for $|a - b\sqrt{d}| < 1$. □

Theorem 5.2.2. *Suppose $d \in \mathbb{N}$ is non-square. Then $x^2 - dy^2 = 1$ has a nontrivial solution, i.e., there is a unit in $\mathbb{Z}[\sqrt{d}]$ of norm 1 other than ± 1.*

Proof. **Step 1.** Fix $B_1 > 1$. Then by Dirichlet's approximation theorem, there exist $a_1, b_1 \in \mathbb{N}$ such that $|a_1 - b_1\sqrt{d}| < \frac{1}{B_1} < \frac{1}{b_1}$. Let $B_2 > b_1$ such that $\frac{1}{B_2} < |a_1 - b_1\sqrt{d}|$. Applying Dirichlet's approximation again, we get a new pair (a_2, b_2) of integers such that

$$|a_2 + b_2\sqrt{d}| < \frac{1}{B_2} < \frac{1}{b_2}.$$

[2]For instance, 0.5006 and 0.49998 are within $\frac{1}{1000}$ of each other, but their first three digits are only equal after rounding to the nearest 4 digits.

Repeating this we see there an infinite sequence of distinct integer pairs (a_j, b_j) such that $|a_j - b_j\sqrt{d}|$ gets smaller and smaller, and

$$|a_j - b_j\sqrt{d}| < \frac{1}{b_j}.$$

for all $j \geq 1$. Then $\frac{a_j}{b_j}$ is an infinite sequence of increasingly good approximations to $\sqrt{d}$.

Step 2. Assume (a, b) satisfies $|a - b\sqrt{d}| < \frac{1}{b}$. Note that

$$|a + b\sqrt{d}| \leq |a - b\sqrt{d}| + |2b\sqrt{d}| \leq 1 + 2b\sqrt{d} \leq 3b\sqrt{d}.$$

Then

$$|a^2 - db^2| = |a + b\sqrt{d}||a - b\sqrt{d}| \leq 3b\sqrt{d}\frac{1}{b} = 3\sqrt{d}.$$

Hence there are infinitely many $a - b\sqrt{d} \in \mathbb{Z}[\sqrt{d}]$ whose norm, in absolute values, is at most $3\sqrt{d}$.

Step 3. By successive applications of the (infinite) pigeonhole principle, we have:
(i) infinitely many $a - b\sqrt{d}$ with the same norm $n \in \mathbb{Z}$, where $|n| \leq 3\sqrt{d}$ (and $n \neq 0$)
(ii) infinitely many $a - b\sqrt{d}$ with norm n and $a \equiv a_0 \bmod n$ for some a_0.
(iii) infinitely many $a - b\sqrt{d}$ with norm n, $a \equiv a_0 \bmod n$, $b \equiv b_0 \bmod n$ for some b_0.

Hence, relabeling if necessary, we have $a_1, b_1, a_2, b_2 \in \mathbb{N}$ such that $N(a_1 - b_1\sqrt{d}) = N(a_2 - b_2\sqrt{d}) = n$, $a_1 \equiv a_2 \bmod n$, $b_1 \equiv b_2 \bmod n$, and $a_1 - b_1\sqrt{d} \neq \pm(a_2 - b_2\sqrt{d})$.

Step 4. Consider

$$\alpha := \frac{a_1 - b_1\sqrt{d}}{a_2 - b_2\sqrt{d}} = \frac{(a_1 - b_1\sqrt{d})(a_2 + b_2\sqrt{d})}{a_2^2 - db_2^2} = \frac{a_1a_2 - db_1b_2}{n} + \frac{a_1b_2 - b_1a_2}{n}\sqrt{d}.$$

Note

$$a_1a_2 - db_1b_2 \equiv a_1a_1 - db_1b_1 \equiv a_1^2 - db_1^2 \equiv 0 \bmod n,$$

and

$$a_1b_2 - b_1a_2 \equiv a_1b_1 - b_1a_1 \equiv 0 \bmod n.$$

Thus the coefficients of α are integers, i.e., $\alpha = a + b\sqrt{d} \in \mathbb{Z}[\sqrt{d}]$ where $a, b \in \mathbb{Z}$. Then hen since

$$a^2 - db^2 = N(a + b\sqrt{d}) = N(a_1 - b_1\sqrt{d})N\left((a_2 - b_2\sqrt{d})^{-1}\right) = nn^{-1} = 1,$$

i.e., (a, b) is a solution of $x^2 - dy^2 = 1$. Furthermore, it is a nontrivial solution since $a_1 - b_1\sqrt{d} \neq \pm(a_2 - b_2\sqrt{d})$. □

Exercise 5.2.1. Explain how to modify the above proof to conclude the existence of infinitely many solutions to $x^2 - dy^2 = 1$. Conclude the real quadratic rings $\mathbb{Z}[\sqrt{d}]$ ($d > 1$ non-square) have infinitely many units, in contrast to the case of imaginary quadratic rings $\mathbb{Z}[\sqrt{-d}]$.

The above approximations suggest how solutions to Pell's equation are related to rational approximations to $\sqrt{d}$.

Proposition 5.2.3. *Suppose (x, y) is a nontrivial solution to $x^2 - dy^2 = 1$. Assume $x, y > 0$. Then*

$$0 < \frac{x}{y} - \sqrt{d} < \frac{1}{y(1+\sqrt{d})} < \frac{1}{2y}.$$

Proof. Let $\alpha = x - y\sqrt{d} \in \mathbb{Z}[\sqrt{d}]$. Then $N(\alpha) = \alpha\overline{\alpha} = 1$. Note $\overline{\alpha} = x + y\sqrt{d} \geq 1 + \sqrt{d}$ since $x, y \geq 1$. Thus $\alpha \leq \frac{1}{1+\sqrt{d}}$, so

$$\frac{x}{y} - \sqrt{d} = \frac{\alpha}{y} < \frac{1}{y(1+\sqrt{d})}.$$

Since $\overline{\alpha}$ and $N(\alpha)$ are positive, so is α, and thus $\frac{\alpha}{y}$, which finishes the asserted bounds. □

This proposition says positive solutions (x, y) to Pell's equation, i.e., units of norm $+1$, give rational approximations to $\sqrt{d}$, and solutions with larger values of y give better approximations. Furthermore, we are always getting overestimates for $\sqrt{d}$. One can similarly get underestimates with units of norm -1, i.e., solutions to $x^2 - dy^2 = -1$, at least when they exist. When they don't, one could instead look for solutions to $x^2 - dy^2 = -2$ or $x^2 - dy^2 = -3$ etc. We remark the approximation in (5.0.1) corresponds to the solution $(30547, 21600)$ to $x^2 - 2y^2 = -791$, though I'm not suggesting Babylonians came up with this approximation by starting with the equation $x^2 - 2y^2 = -791$!

Exercise 5.2.2. Suppose (x, y) is solution to $x^2 - dy^2 = -1$ with $x, y > 0$. Show $\frac{x}{y} < \sqrt{d}$ and prove a (good) bound for $\sqrt{d} - \frac{x}{y}$ in terms of y.

Exercise 5.2.3. Suppose (x, y) is solution to $x^2 - dy^2 = -2$ with $x, y > 0$. Show $\frac{x}{y} < \sqrt{d}$ and prove a (good) bound for $\sqrt{d} - \frac{x}{y}$ in terms of y.

Example 5.2.1. Recall from Example 5.1.1, $(3, 2)$, $(17, 12)$ and $(99, 70)$ are solutions to $x^2 - 2y^2 = 1$. This gives the following approximations, with the following error bounds $\frac{1}{2y}$ from the above proposition:

$\frac{x}{y}$	decimal	error bound	$\frac{x}{y} - \sqrt{2}$
$\frac{3}{2}$	1.5	< 0.25	$0.085786...$
$\frac{17}{12}$	$1.41\overline{6}$	$< 0.041\overline{6}$	$0.0024531...$
$\frac{99}{70}$	$1.\overline{4142857}$	$< 0.00\overline{714285}$	$0.00007215...$

Exercise 5.2.4. Recall the solutions in the previous example came from the first three powers ϵ^n $(n = 1, 2, 3)$ of the unit $\epsilon = 3 - 2\sqrt{2} \in U_2^+$. However, none of these approximations are better than the (more complicated) Babylonian one (5.0.1). Using a calculator, compute a few successive approximations to $\sqrt{2}$ from higher powers ϵ^n, along with the exact error (up to several decimal places). What is the first approximation you get in this way that is better (i.e., closer to $\sqrt{2}$) than the one in (5.0.1).

Exercise 5.2.5. Using 3 nontrivial solutions to $x^2 - 3y^2 = 1$ you found in Exercise 5.1.3, give 3 rational approximations to $\sqrt{3}$ with error bounds. Using a calculator, compute the actual error in these approximations (e.g., you can make a table as in the example above).

5.3 Fundamental units

Here we determine the structure of the group of units U_d, which will give us a method for generating all solutions to Pell's equation.

Definition 5.3.1. *The* **fundamental unit** ε_d *of* $\mathbb{Z}[\sqrt{d}]$ *is the smallest unit* $x+y\sqrt{d} \in \mathbb{Z}[\sqrt{d}]$ *such that* $x, y > 0$. *The* **fundamental +unit** ε_d^+ *of* $\mathbb{Z}[\sqrt{d}]$ *is the smallest unit* $x + y\sqrt{d} \in \mathbb{Z}[\sqrt{d}]$ *such that* $x, y > 0$ *and* $N(\epsilon) = 1$.[3]

In real quadratic rings, smallest means with respect to the usual order on $\mathbb{R}$, unlike the case of imaginary quadratic rings where we (partially) ordered elements by their norm.

Lemma 5.3.2. *For any (non-square)* $d > 1$, *the fundamental unit* ε_d *and the fundamental +unit* ε_d^+ *exist and are uniquely defined.*

Proof. Since $<$ defines a strict ordering of real numbers, the condition of "smallest" guarantees that ε_d and ε_d^+ will be unique if they exist, so it suffices to show existence.

Recall we always have a nontrivial solution (x_0, y_0) to $|x^2 - dy^2| = 1$ from Theorem 5.2.2. Moreover, we can assume $x_0, y_0 > 0$. Now note if $x + y\sqrt{d} < \varepsilon_0 = x_0 + y_0\sqrt{d}$ with $x, y > 0$, we must have $x < \varepsilon_0$ and $y < \frac{\varepsilon_0}{\sqrt{d}}$. Hence

$$\varepsilon_d = \min\left\{x + y\sqrt{d} : 1 \leq x, \sqrt{d}y < \varepsilon_0,\ |x^2 + dy^2| = 1\right\}.$$

Since the set on the right is finite, this minimum is well defined, hence ε_d exists.

The case of ε_d^+ follows in the same way, simply using the equation $x^2 - dy^2 = 1$ instead of $|x^2 - dy^2| = 1$. □

[3]Note that most discussions you will find about fundamental units talk about fundamental units in the ring of integers $\mathcal{O}_d$ of $\mathbb{Q}(\sqrt{d})$. Here, assuming d is squarefree, $\mathcal{O}_d$ is just $\mathbb{Z}[\sqrt{d}]$ when $d \not\equiv 1 \bmod 4$ but is $\mathbb{Z}[\frac{1+\sqrt{d}}{2}]$ when $d \equiv 1 \bmod 4$ (recall Definition 2.5.5). So be careful comparing what we say here and what is written other places about fundamental units, as there may be a slight difference when $d \equiv 1 \bmod 4$, though there is no serious difference in the theory. E.g., when $d = 5$ a fundamental unit in $\mathbb{Z}[\sqrt{5}]$ is $2 + \sqrt{5}$ but in $\mathcal{O}_5$ it is $\frac{1+\sqrt{5}}{2}$. On the other hand, the fundamental unit in $\mathcal{O}_{17} = \mathbb{Z}[\frac{1+\sqrt{17}}{2}]$ is the same as the fundamental unit in $\mathbb{Z}[\sqrt{17}]$, namely $\varepsilon_{17} = 4 + \sqrt{17}$.

Also, the term "fundamental +unit" is not standard—as far as I know, there is no standard term for the phrase "the smallest unit of norm 1."

Here is a naive algorithm for finding ε_d or ε_d^+. First, pick some bound $N \geq 1$. Then range over all $1 \leq x, y \leq N$ to look for solutions to $|x^2 - dy^2| = 1$ or $x^2 - dy^2 = 1$. If we found any, then the smallest solution (x, y) gives us ε_d or ε_d^+ as $x + y\sqrt{d}$. If not, we pick a larger N and repeat. This process terminates at some point by the existence of ε_d and ε_d^+.

Example 5.3.1. When $d = 2$, $\varepsilon_2 = 1 + \sqrt{2}$ and $\varepsilon_2^+ = \varepsilon_2^2 = 3 + 2\sqrt{2}$. We saw both of these units in Example 5.1.1.

Example 5.3.2. When $d = 5$, we compute $\varepsilon_5 = 2 + \sqrt{5}$ and $\varepsilon_5^+ = \varepsilon_5^2 = 9 + 4\sqrt{5}$.

Example 5.3.3. Consider $d = 7$. Recall from Exercise 5.1.4 that $\mathbb{Z}[\sqrt{7}]$ has no units of norm -1. We compute $\varepsilon_7 = \varepsilon_7^+ = 8 + 3\sqrt{7}$.

Exercise 5.3.1. Compute ε_d and ε_d^+ for $d = 3, 6, 11$.

Exercise 5.3.2. An alternative definition of fundamental unit (resp. +unit) is the smallest $\varepsilon > 1$ in $\mathbb{Z}[\sqrt{d}]$ such that $|N(\varepsilon)| = 1$) (resp. $N(\epsilon) = 1$). Prove that this is equivalent to the above definition as follows. (*Suggestion:* Show that $\varepsilon = x + y\sqrt{d} > 1$ a unit implies $|\overline{\varepsilon}| < 1$, which implies $x, y > 0$.)

Theorem 5.3.3. *For $d > 1$ non-square, U_d (resp. U_d^+) is the infinite abelian group generated by ε_d (resp. ε_d^+) and -1. Explicitly,*

$$U_d = \{\ldots, \pm\varepsilon_d^{-2}, \pm\varepsilon_d^{-1}, \pm 1, \pm\varepsilon_d, \pm\varepsilon_d^2, \ldots\}$$

and

$$U_d^+ = \{\ldots, \pm(\varepsilon_d^+)^{-2}, \pm(\varepsilon_d^+)^{-1}, \pm 1, \pm\varepsilon_d^+, \pm(\varepsilon_d^+)^2, \ldots\},$$

and all the elements listed in the sets on the right are distinct, i.e., $\varepsilon_d^m = \pm\varepsilon_d^n$ for $m, n \in \mathbb{Z}$ (resp. $(\varepsilon_d^+)^m = \pm(\varepsilon_d^+)^n$) if and only if $m = n$ and the plus/minus sign is $+$.

Proof. We know U_d and U_d^+ are abelian groups by Proposition 3.3.4 and Exercise 5.1.1. Thus U_d and U_d^+ must contain all the elements in the sets on the right.

Write G denote U_d or U_d^+, and let ε denote ε_d or ε_d^+, according to whether $G = U_d$ or $G = U_d^+$. Since $\varepsilon > 1$, the sequence ε^n $(n \geq 0)$ is a strictly increasing sequence lying in $[1, \infty)$, and ε^{-n} $(n > 0)$ is a strictly decreasing sequence lying in $(0, 1)$. From this one easily sees that all elements in the above sets on the right are distinct.

Finally, we show any $\alpha \in G \subset \mathbb{Z}[\sqrt{d}]$ is of the form $\pm\varepsilon^n$ for some $n \in \mathbb{Z}$. Suppose there is some α which is not of this form. By taking the negative and/or inverse if necessary, we may assume $\alpha > 1$. Since ε is the smallest element of G larger than 1 (Exercise 5.3.2) and $\varepsilon^n \to \infty$ as $n \to \infty$, there must be some $n > 0$ such that $\varepsilon^m < \alpha < \varepsilon^{m+1}$. But then $1 < \alpha\varepsilon^{-m} < \varepsilon$ and $N(\alpha\varepsilon^{-m}) = 1$, contradicting the minimality of ε. □

Note that for any unit $u \in U_d$, $N(u) = u\overline{u} = \pm 1$ implies u^{-1} is either $\overline{u}$ or $-\overline{u}$, according to whether $N(u) = 1$ or $N(u) = -1$. In particular, if $(\varepsilon_d^+)^n = x_n + y_n\sqrt{d}$, then $(\varepsilon_d^+)^{-n} = \overline{(\varepsilon_d^+)^n} = x_n - y_n\sqrt{d}$.

Hence the above theorem says that once we find ε_d^+, we can compute all elements of U_d^+ by computing $(\varepsilon_d^+)^n = x_n + y_n\sqrt{d}$, $n \geq 1$. Then the elements of U_d^+ are

$$U_d^+ = \{\pm 1\} \cup \left\{ \pm x_n \pm y_n\sqrt{d} : n \geq 1 \right\}, \tag{5.3.1}$$

where we read the $\pm$ signs in $\pm x_n \pm y_n$ independently. (A similar statement is also true for U_d.) This immediately gives our desired description of solutions to (5.0.2).

Corollary 5.3.4. *For $n \geq 1$, write $(\varepsilon_d^+)^n = x_n + y_n\sqrt{d}$ for $n \geq 1$ (with $x_n, y_n \in \mathbb{Z}$). Then all solutions to Pell's equation $x^2 - dy^2 = 1$ are the trivial solutions $(\pm 1, 0)$ and the nontrivial solutions $(\pm x_n, \pm y_n)$ for $n \geq 1$.*

Via Proposition 5.2.3, this gives us the following sequence of approximations

$$\frac{x_n}{y_n} \approx \sqrt{d}$$

of $\sqrt{d}$. To prove that these approximations are getting better (at least asymptotically), by this proposition we want to prove the y_n's are increasing.

Exercise 5.3.3. With x_n, y_n as above, show the sequences (x_n) and (y_n) are strictly increasing sequences for $n \geq 1$. Deduce that the sequence $\frac{x_n}{y_n}$ converges to $\sqrt{d}$.

Using the above theorem, we can also relate ε_d and ε_d^+ now.

Exercise 5.3.4. For $d > 1$ a non-square, show $\varepsilon_d^+ = \varepsilon_d^2$ if $\mathbb{Z}[\sqrt{d}]$ has units of norm -1, and $\varepsilon_d^+ = \varepsilon_d$ otherwise. Deduce in particular that $\varepsilon_d \leq \varepsilon_d^+$.

Hence if we solve the problem of finding the fundamental unit ε_d, we also know the fundamental +unit ε_d^+. Since $\varepsilon_d \leq \varepsilon_d^+$, even if our goal is to compute ε_d^+, it may often be easier algorithmically to look for ε_d first, since the x and y appearing in the representation $x + y\sqrt{d}$ can be much smaller.

Example 5.3.4. Consider $d = 29$. Then by the naive algorithm for finding fundamental units, we can check $\varepsilon_{29} = 70 + 13\sqrt{29}$, which has norm -1. Thus $\varepsilon_{29}^+ = \varepsilon_{29}^2 = 9801 + 1820\sqrt{29}$, but this would require many more calculations to find solely by the naive algorithm.

Here's another consequence of the structure theorem for U_d (or, if you prefer, the previous exercise): if $\mathbb{Z}[\sqrt{d}]$ has no units of norm -1, we can prove this algorithmically by computing ε_d and checking it has norm 1. For then $\pm\varepsilon_d^n$ also has norm 1 for all n, i.e., $U_d = U_d^+$.

5.4 Continued fractions

In the last section, we described how to find all solutions to Pell's equation in terms of the fundamental +unit ε_d^+. Earlier, we also presented a naive algorithm to compute ε_d and ε_d^+. The problem is that as d gets even moderately large, the naive algorithm is not very efficient. This is already suggested by the case of $d = 29$ in Example 5.3.4. Here is a more impressive example:

Example 5.4.1. When $d = 61$, $\varepsilon_d^+ = 1766319049 + 226153980\sqrt{61}$, i.e., the smallest positive nontrivial solution to $x^2 - 61y^2 = 1$ is $(1766319049, 226153980)$.

The above example was discovered by Bhaskara II in the 12th century in India, and independently (much later) rediscovered by Fermat in Europe. How can one find such solutions, especially without powerful computing devices? The answer come from an alternative representation of numbers, not as decimals, but as *continued fractions*.

First we explain the continued fraction expansion with an example.

Example 5.4.2. Consider $\frac{a}{b} = \frac{a_1}{b_1} = \frac{13}{5}$, which we write as a whole number plus a remainder:

$$\frac{a_1}{b_1} = \frac{13}{5} = 2 + \frac{3}{5}.$$

Now we can't exactly repeat this on the remainder, but we can on its *reciprocal*:

$$\frac{a_2}{b_2} := \frac{5}{3} = 1 + \frac{2}{3}.$$

Thus we have

$$\frac{a}{b} = 2 + \frac{1}{3/2} = 2 + \frac{1}{1 + \frac{2}{3}}.$$

Now repeat again with the reciprocal of the remainder in $\frac{a_2}{b_2}$:

$$\frac{a_3}{b_3} := \frac{3}{2} = 1 + \frac{1}{2}.$$

If we do this again, we get:

$$\frac{a_4}{b_4} = \frac{2}{1} = 2,$$

a rational with no remainder, and so we stop. This leads to the following expression:

$$\frac{a}{b} = \frac{13}{5} = \boxed{2} + \cfrac{1}{\boxed{1} + \cfrac{1}{\boxed{1} + \cfrac{1}{\boxed{2}}}},$$

which we call the continued fraction expansion of $\frac{13}{5}$. Note that because at each stage we are taking reciprocals, we'll see a sequence of 1's going down, and all that matters are the

numbers in the boxes. To simplify notation, we will also write this as

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

Definition 5.4.1. *Let $x \in \mathbb{R}$. The* **continued fraction expansion** *of x is the expression $[q_1, q_2, q_3, \ldots] = q_1 + \frac{1}{q_2+}\frac{1}{q_3+}\cdots$ where $q_1 \in \mathbb{Z}$ and $q_j \in \mathbb{Z}_{\geq 0}$ for $j \geq 2$ are defined as follows:*

- *$q_1 = \lfloor x \rfloor$ is the greatest integer $\leq x$, so $0 \leq r_1 < 1$ where $r_1 = x - q_1$;*
- *for $j \geq 1$, inductively set*

$$q_{j+1} = \begin{cases} \lfloor \frac{1}{r_j} \rfloor \text{(the greatest integer} \leq r_j) & r_j \neq 0 \\ 0 & r_j = 0, \end{cases}$$

so $r_{j+1} := r_j - q_j$ satisfies $0 \leq r_{j+1} < 1$.

If $r_j = 0$ for all $j > m$, we also write the continued fraction expansion as the finite sequence $[q_1, \ldots, q_m] = q_1 + \frac{1}{q_2+}\frac{1}{q_3+}\cdots\frac{1}{q_m}$, in which case we call the continued fraction expansion **finite**.

The q_j and r_j is used to make you think that these quantities are like quotients and remainders (which they are if x is rational). The rounding down function $x \mapsto \lfloor x \rfloor$ (also often denoted by $x \mapsto [x]$) is called the **greatest integer function** or the **floor function**.

Note for any $x \in \mathbb{R}$, there is a unique continued fraction expansion $[q_1, q_2, \ldots]$. Moreover, since at each step $0 \leq r_j < 1$, the reciprocal will be at least 1 if $r_j \neq 0$, and so $q_{j+1} = 0$ if and only if $r_j = 0$.

Exercise 5.4.1. Compute the continued fraction expansion of $\frac{80}{17}$.

Exercise 5.4.2. Let $x \in \mathbb{R}$, and $[q_1, q_2, \ldots]$ be the continued fraction expansion. Let (x_n) denote the sequence of rational numbers by evaluation the partial continued fraction expansions:

$$x_n = q_1 + \cfrac{1}{q_2 + \cfrac{1}{\ddots_{+\frac{1}{q_n}}}}.$$

Show $\lim_{n\to\infty} x_n = x$.

Exercise 5.4.3. For $x \in \mathbb{R}$, show the continued fraction expansion for x is finite if and only if $x \in \mathbb{Q}$.

Example 5.4.3. Let's compute the continued fraction expansion $[q_1, q_2, \ldots]$ of $\sqrt{5}$.

First set

$$q_1 = \lfloor \sqrt{5} \rfloor = 2, \quad r_1 = \sqrt{5} - 2,$$

so at the first stage our expansion looks like

$$\sqrt{5} = q_1 + r_1 = 2 + (\sqrt{5} - 2) = 2 + \frac{1}{1/(\sqrt{5} - 2)}.$$

The nice thing about quadratic numbers is we can rationalize the denominator in $\frac{1}{\sqrt{5}-2}$ by multiplying by the conjugate (in $\mathbb{Z}[\sqrt{5}]$) of the denominator. Note $N(r_1) = r_1 \overline{r_1} = N(-2 + \sqrt{5}) = 4 - 5 = -1$, so $\frac{1}{r_1} = -\overline{r_1} = 2 + \sqrt{5}$, i.e.,

$$\frac{1}{r_1} = \frac{1}{\sqrt{5} - 2} = 2 + \sqrt{5}.$$

So at the next stage we let

$$q_2 = \lfloor 2 + \sqrt{5} \rfloor = 4, \quad r_2 = 2 + \sqrt{5} - q_2 = \sqrt{5} - 2 = r_1.$$

Thus at the next stage, we have the expansion

$$\sqrt{5} = 2 + \frac{1}{4 + \frac{1}{\sqrt{5}-2}}.$$

Since $r_2 = r_1$, we see that $q_3 = q_2$, so $r_3 = r_2 = r_1$, and so on. So these computations simply repeat, and we will have $q_j = 4$ for all $j \geq 2$, giving the continued fraction expansion

$$\sqrt{5} = [2, 4, 4, 4, \ldots] = 2 + \frac{1}{4+}\frac{1}{4+}\frac{1}{4+} \cdots .$$

Example 5.4.4. Now let's try finding the continued fraction expansion of $\sqrt{3}$. At the first stage we have

$$q_1 = \lfloor \sqrt{3} \rfloor = 1, \quad r_1 = \sqrt{3} - 1.$$

Then

$$\frac{1}{r_1} = \frac{1}{\sqrt{3} - 1} \frac{\sqrt{3} + 1}{\sqrt{3} + 1} = \frac{1 + \sqrt{3}}{2}.$$

So

$$q_2 = \lfloor \frac{1 + \sqrt{3}}{2} \rfloor = 1, \quad r_2 = \frac{1 + \sqrt{3}}{2} - 1 = \frac{\sqrt{3} - 1}{2}.$$

Then

$$\frac{1}{r_2} = \frac{2}{\sqrt{3} - 1} \frac{\sqrt{3} + 1}{\sqrt{3} + 1} = 1 + \sqrt{3}.$$

So

$$q_3 = \lfloor 1 + \sqrt{3} \rfloor = 2, \quad r_3 = \sqrt{3} - 1 = r_1.$$

Since $r_3 = r_1$, we must repeat after this, i.e., $q_4 = q_2$, $r_4 = r_2$, and so on, giving us the continued fraction expansion

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

Notice the above continued fraction expansions repeat. We make this notion precise as follows.

Definition 5.4.2. *We say a continued fraction expansion $[q_1, q_2, \ldots]$ is* **periodic** *if there exist $s \geq 0$ and $m \in \mathbb{N}$ such that*

$$[q_1, q_2, \ldots] = [q_1, \ldots, q_s, q_{s+1}, \ldots, q_{s+m}, q_{s+1}, \ldots, q_{s+m}, q_{s+1}, \ldots, q_{s+m}, \ldots],$$

i.e., if $q_{j+m} = q_j$ for all $j > s$. In this case, we denote this expansion by

$$[q_1, \ldots, q_s, \overline{q_{s+1}, \cdots, q_{s+m}}].$$

If the expansion is periodic, the smallest such m for which the above condition holds (for some s) is called the **period** *of $[q_1, q_2, \ldots]$.*

For instance, the examples above say $\sqrt{5} = [2, \overline{4}]$ is periodic with period 1 and $\sqrt{3} = [1, \overline{1, 2}]$ is periodic with period 2. Since any rational number has continued fraction expansion of the form $[q_1, \ldots, q_s, \overline{0}]$, any rational number has a periodic continued fraction expansion with period 1. Note periodic continued fractions can be specified by a finite amount of data.

Theorem 5.4.3 (Lagrange)**.** *For any $x \in \mathbb{R}$, the continued fraction expansion of x is periodic if and only if $x \in \mathbb{Q}(\sqrt{d})$ for some $d \geq 1$. In particular the continued fraction expansion of any element of $\mathbb{Z}[\sqrt{d}]$ is periodic.*

Due to lack of time, we won't prove this. But the idea of the proof for the "if" direction is that at each stage in the continued fraction expansion, the quantities $\frac{1}{r_j}$ will be elements of $\mathbb{Q}(\sqrt{d})$ satisfying certain conditions, and then showing that there are only finitely many possibilities, so for some $j, m \geq 1$, we have $r_{j+m} = r_j$ by the pigeonhole principle.

The "only if" direction is easier. For simplicity, we just illustrate the special case $s = 0$, so $x = [\overline{q_1, \ldots, q_m}]$ (what is called *purely periodic*). Consider $\alpha = [q_1, \ldots, q_m] \in \mathbb{Q}$. Then

$$x = \alpha + \frac{1}{\alpha + \frac{1}{\alpha + \cdots}}.$$

Then

$$x - \alpha = \frac{1}{\alpha + \frac{1}{\alpha + \cdots}},$$

so taking reciprocals shows

$$\frac{1}{x - \alpha} = \alpha + \frac{1}{\alpha + \frac{1}{\alpha + \cdots}} = x,$$

i.e.,

$$1 = (x - \alpha)x = x^2 - \alpha x,$$

so x^2 satisfies the quadratic equation $x^2 - \alpha x - 1 = 0$ with rational coefficients, from which it follows $x \in \mathbb{Q}(\sqrt{d})$ where $d = \alpha^2 + 4$.

Exercise 5.4.4. Compute the continued fraction expansion of $\sqrt{2}$.

Exercise 5.4.5. Compute the continued fraction expansion of $\sqrt{7}$.

Now we explain (without proof) the connection with Pell's equation and fundamental units. Assume $d > 1$ is squarefree, and write

$$d = [q_1, \dots, q_s, \overline{q_{s+1}, \dots, q_{s+m}}]$$

where s and m are chosen minimally so we can represented this continued fraction periodically. In particular m is the period. Consider the partial continued fraction expansions $[q_1, \dots, q_n]$. These are rational numbers, so we write them as

$$\frac{x_n}{y_n} = [q_1, \dots, q_n], \quad x_n, y_n \in \mathbb{N}, \gcd(x_n, y_n) = 1.$$

As they converge to $\sqrt{d}$ (Exercise 5.4.2), we call (x_n, y_n) the n-th **convergent** of the continued fraction.

Theorem 5.4.4. *For $d > 1$ squarefree, let (x_n, y_n) be the n-th convergent in the continued fraction expansion of $\sqrt{d}$. Then $x_m + y_m\sqrt{d}$ is the fundamental unit ε_d, where m is the period of this continued fraction. More generally, $x_{km} + y_{km}\sqrt{d}$ is ε_d^k for $k \geq 1$.*

So in summary, we used units in real quadratic fields to determine all solutions to Pell's equation in terms of ε_d. Now we know how to compute ε_d in terms of continued fractions, and thus determine all solutions to Pell's equation. Moreover, this allows us to construct good rational approximations to $\sqrt{d}$. Of course, using continued fractions directly gives us rational approximations to $\sqrt{d}$, but in some sense the ones coming from solutions to Pell's equation (or $x^2 - dy^2 = -1$) are optimal in that they will have minimal remainder (see Proposition 5.2.3). (We also haven't proved that the continued fraction convergents give us good rational approximations, though one can prove this.)

Example 5.4.5. Recall $\sqrt{5} = [2, \overline{4}]$, which has period 1. So we look at the first convergent is given by

$$\frac{x_1}{y_1} = [q_1] = [2] = \frac{2}{1}$$

so the theorem says $\varepsilon_5 = 2 + \sqrt{5}$, which matches with Example 5.3.2.

Example 5.4.6. Recall $\sqrt{2} = [1, \overline{1, 2}]$, which has period 2. Thus the second convergent is given by

$$\frac{x_2}{y_2} = [1, 1] = 1 + \frac{1}{1} = \frac{2}{1},$$

so the theorem says $\varepsilon_3 = 2+\sqrt{3}$, which matches what you should have got in Exercise 5.3.1.

Exercise 5.4.6. Use the continued fraction expansion of $\sqrt{7}$ to compute ε_7 (see also Example 5.3.3). Then obtain a rational approximation to $\sqrt{7}$ accurate to within $\frac{1}{100}$.

Exercise 5.4.7. Use the continued fraction expansion of $\sqrt{19}$ to compute ε_{19}. Use ε_{19} to obtain a rational approximation to $\sqrt{19}$ accurate to within $\frac{1}{1000}$. (You may use a calculator.)

Exercise 5.4.8. Use continued fractions to obtain the expression for ε_{61}^+ asserted in Example 5.4.1. (You may use a calculator.)

5.5 Aftermission: fundamental units and Fibonacci numbers

We close this chapter with an amusing connection with fundamental units and Fibonacci numbers, following ideas that we used to solve Pell's equation.

The **golden ratio** $\phi = \frac{1+\sqrt{5}}{2}$ is the fundamental unit for the full ring of integers $\mathbb{Z}[\frac{1+\sqrt{5}}{2}]$. For $x, y \in \mathbb{Z}$, note

$$N(x + y\frac{1+\sqrt{5}}{2}) = (x + y\frac{1+\sqrt{5}}{2})(x + y\frac{1-\sqrt{5}}{2}) = x^2 + xy - y^2.$$

This expression is a binary quadratic form, which we also denote

$$Q(x, y) = x^2 + xy - y^2.$$

Recall the **Fibonacci numbers** F_n are defined by

$$F_1 = F_2 = 1, F_{n+2} = F_{n+1} + F_n, \quad n \geq 1.$$

Exercise 5.5.1. Show the Fibonacci numbers satisfy $F_{2n+2}^2 + 1 = F_{2n+2}F_{2n+1} + F_{2n+1}^2$.

Put another way, the exercise says that (F_{2n+1}, F_{2n+2}) are solutions to

$$Q(x, y) = 1,$$

which is an analogous equation to Pell's equation. In other words

$$F_{2n+1} + F_{2n+2}\frac{1+\sqrt{5}}{2}$$

is a unit of norm 1 in $\mathbb{Z}[\frac{1+\sqrt{5}}{2}]$.

Note the golden ratio ϕ has norm -1, but it square $\varepsilon = \phi^2 = \frac{3+\sqrt{5}}{2}$ has norm 1 (i.e., is the fundamental +unit in $\mathbb{Z}[\frac{1+\sqrt{5}}{2}]$. Then we can write

$$\varepsilon = \frac{3+\sqrt{5}}{2} = 1 + 1 \cdot \frac{1+\sqrt{5}}{2} = F_1 + F_2 \frac{1+\sqrt{5}}{2}.$$

Computing a couple of powers of ε, we see

$$\varepsilon^2 = \frac{7+3\sqrt{5}}{2} = 2 + 3 \cdot \frac{1+\sqrt{5}}{2} = F_3 + F_4 \frac{1+\sqrt{5}}{2},$$

$$\varepsilon^3 = \frac{47+21\sqrt{5}}{2} = 13 + 21 \cdot \frac{1+\sqrt{5}}{2} = F_5 + F_6 \frac{1+\sqrt{5}}{2}.$$

This is part of a general rule.

Exercise 5.5.2. Compute ε^4 directly and then check that $\varepsilon_4 = F_7 + F_8 \frac{1+\sqrt{5}}{2}$.

Exercise 5.5.3. Prove that $\varepsilon^n = F_{2n-1} + F_{2n}\phi$ for $n \geq 1$.

The above expression gives a way to compute Fibonacci numbers. While it's not exactly presented as a formula for F_n, you probably noticed in the calculations above you immediately see F_{2n} as the coefficient b in the expression $\varepsilon^n = \frac{a+b\sqrt{5}}{2}$, and then a is just $b + 2F_{2n-1}$. One can rewrite these calculations into a well-known formula for F_n:

Exercise 5.5.4. Prove that $F_n = \frac{\phi^n - \overline{\phi}^n}{\phi - \overline{\phi}}$ for $n \geq 1$.

Chapter 6

The Last Theorem

NOTE: This chapter is just an outline as we didn't have time to cover this this semester.

More generally that just getting non-existence of solutions to certain Diophantine equations, we can also obtain necessary conditions for solutions to Diophantine equations. We illustrate this here by saying a little bit about Fermat's Last Theorem.

Let $n \in \mathbb{N}$. A solution (x, y, z) to the Diophantine equation

$$x^n + y^n = z^n \tag{6.0.1}$$

is said to be **trivial** if $xyz = 0$, i.e., if at least one of x, y and z is 0. Note there are infinitely many solutions with $xyz = 0$ and they are easy to describe. E.g., if $y = 0$, this reduces to $x^n = z^n$, which means $x = z$ or $x = \pm z$, depending on whether n is odd or even.

Theorem 6.0.1. (Fermat's last theorem (FLT)) *For $n \geq 3$, $x^n + y^n = z^n$ has no nontrivial solutions over $\mathbb{Z}$. In particular, it has no solutions in positive integers.*

The story is Fermat claimed in 1637 in a margin of a copy of *Arithmetica* (an Ancient Greek text by Diophantus) to have a beautiful proof, but said the proof could not fit in the margin. For centuries, mathematicians tried to find a proof of this, and it was eventually proven in 1995 by Andrew Wiles with help from Richard Taylor, building on the work of many others and using mathematics far beyond what was available in Fermat's time. For a long time, people wondered if Fermat really had a proof, and this was one of the romanticized mysteries of mathematics. But what is generally suspected now is that Fermat did have proofs for $n = 3$ and $n = 4$, and possibly some other cases, and probably he thought he had an argument which would work in general but turned out to be incorrect (which Fermat may or may not have realized himself later).

I had hoped to have 2–3 lectures to discuss Fermat's last theorem in class, but ran out of time. Here is what I had hoped to do:

- reduce proving Fermat's last theorem to the cases $n = 4$ and n is prime (it's an easy exercise you can do yourself)
- prove the $n = 4$ case of Fermat's last theorem using descent like Fermat, by showing that $x^4 - y^4 = z^2$ has no "primitive" solutions

- prove or at least sketch the $n = 3$ case of Fermat's last theorem using unique factorization in $\mathbb{Z}[\zeta_3]$
- explain, roughly, how knowing unique factorization in $\mathbb{Z}[\zeta_p]$ would prove the $n = p$ case of Fermat's last theorem (in 1847 Lamé used this idea to give a flawed proof—Kummer noted the proof doesn't work for all primes because he already knew $\mathbb{Z}[\zeta_p]$ doesn't always have unique factorization[1]), and what we know about when $\mathbb{Z}[\zeta_p]$ has unique factorization or not

Maybe I will write some of this the next time I teach this course. I only include now a partial elementary result on the $n = 3$ case of Fermat's Last Theorem.

Proposition 6.0.2. *If there is a nontrivial solution to $x^3 + y^3 = z^3$ (i.e., a solution over $\mathbb{Z}$ with x, y, z all nonzero), then either exactly one of x, y, z divisible is by 7 or all of x, y, z are divisible by 7.*

Proof. First, if there is a nontrivial solution to $x^3+y^3 = z^3$, we may replace x, y and z with x/d, y/d and z/d where d is the gcd of x, y and z to get what is called a primitive solution, i.e., one where no prime p divides all of x, y and z. So assume (x, y, z) is a primitive solution. In particular, not all of x, y, z are divisible by 7. We want to show exactly one of these is.

Note the cubes mod 7 are 0, $1 \equiv 1^3 \equiv 2^3 \equiv 4^3 \bmod 7$ and $-1 \equiv 6 \equiv 3^3 \equiv 5^3 \equiv 6^3 \bmod 7$. Then we have $x^3 + y^3 \equiv z^3 \bmod 7$ where $x^3, y^3, z^3 \equiv 0, \pm 1 \bmod 7$. It is easy to see this is only possible if exactly one of $x^3, y^3, -z^3$ is 1 mod 7, exactly one is -1 mod 7 and exactly one is 0 mod 7. □

In other words, while modular arithmetic does not allow us to clearly rule out non-trivial solutions to $x^3 + y^3 = z^3$ (at least mod 7), it puts some constraints on non-trivial solutions. Namely, it says at least one of x, y, z must be divisible by 7. One might hope that by putting enough constraints on non-trivial solutions one can prove no non-trivial solutions exist. The reason that looking mod 7 works is that there aren't too many cubes mod 7. Using some basic group theory, one can show that the number of non-zero cubes mod p is $\frac{p-1}{3}$ if $p \equiv 1 \bmod 3$ and $p - 1$ otherwise. (Something similar is true mod m where one looks at whether $3|\phi(m)$ or not.) Thus if we want to try to push this idea further, we should look mod primes p which are 1 mod 3. The next case, $p = 13$, is an exercise.

Exercise 6.0.1. Prove that a non-trivial solution to $x^3 + y^3 = z^3$ must have at least of x, y, z divisible by $p = 13$.

One can also show the analogous statement for $p = 19$. If one could prove such a statement holds for any prime $p \equiv 1 \bmod 3$ (there are infinitely many such primes by a theorem of Dirichlet to be mentioned in the next chapter), then this would mean that (by the infinite pigeonhole principle) for a non-trivial solution to $x^3 + y^3 = z^3$, at least one of x, y, z must be divisible by infinitely primes, which is impossible. This would give a "modular arithmetic" proof (albeit a complicated one) of the $n = 3$ case of FLT. However, analogous

[1] So now Lamé is probably best known in number theory for this mistake. *C'est dommage!* He seems to otherwise have been a good mathematician.

statements are not true for all primes $p \equiv 1 \bmod 3$: $x^3 + y^3 \equiv z^3 \bmod 31$ has solutions with x, y, z all non-zero mod 31. This suggests that one really does need a more sophisticated approach to prove FLT.

Chapter 7

Riemann Zeta Function

NOTE: This chapter is just an outline as we didn't have time to cover this this semester.

In 1859, Bernhard Riemann utterly transformed analytic number theory with a 10-page paper on what is now known as the *Riemann zeta function*, his only work in number theory.[1] Here is roughly what I hoped to say about it:

- Explain the definition $\zeta(s) = \sum \frac{1}{n^s}$, which makes sense for complex numbers $s = r + it$ and converges when $r > 1$, but can be extended to an analytic function for $s \neq 1$.
- Explain the *Euler product* $\zeta(s) = \prod \frac{1}{1-p^{-s}}$ (again converging for $r > 1$) and how this is equivalent to the fundamental theorem of arithmetic.
- Explain in more detail Euler's proof of the infinitude of primes, discussed at the end of the introduction, which using the fact that

$$\zeta(1) = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots = \infty.$$

- Explain the *prime number theory*, which tells us about the distribution of primes, and how this is related to $\zeta(s)$. Also explain the relation with the conjectural *Riemann hypothesis*, about where $\zeta(s) = 0$, which is the most famous open problem in number theory now (after the fall of Fermat's last theorem).
- Make some comments about on one hand primes seem to be distributed randomly, but statistically they obey very precise arithmetic laws (e.g., there are asymptotically the same number of primes 1 mod 4 as there are 3 mod 4).

[1] I hope to have a similar impact someday on the world of mathematical comedy. My imagined eulogy: *He invented the 3-minute riff on Pell's equation, and his theoretical work establishing a ring structure on stand-up jokes is now used by lecturers everywhere. He died the way he lived, stabbing by students.*

Bibliography

[Cox13] David A. Cox, *Primes of the form $x^2 + ny^2$*, second edition ed., Pure and Applied Mathematics (Hoboken), John Wiley & Sons, Inc., Hoboken, NJ, 2013, Fermat, class field theory, and complex multiplication. MR 3236783

[CS03] John H. Conway and Derek A. Smith, *On quaternions and octonions: their geometry, arithmetic, and symmetry*, A K Peters, Ltd., Natick, MA, 2003. MR 1957212 (2004a:17002)

[FR15] Sylvia Forman and Agnes M. Rash, *The whole truth about whole numbers*, Springer, Cham, 2015, An elementary introduction to number theory. MR 3309165

[JJ98] Gareth A. Jones and J. Mary Jones, *Elementary number theory*, Springer Undergraduate Mathematics Series, Springer-Verlag London, Ltd., London, 1998. MR 1610533

[Mara] Kimball Martin, *Number Theory I course notes (Fall 2009)*, http://www.math.ou.edu/~kmartin/nti/.

[Marb] ———, *Number Theory II course notes (Spring 2010)*, http://www.math.ou.edu/~kmartin/ntii/.

[Rou91] G. Rousseau, *On the quadratic reciprocity law*, J. Austral. Math. Soc. Ser. A **51** (1991), no. 3, 423–425. MR 1125443

[Ste09] William Stein, *Elementary number theory: primes, congruences, and secrets*, Undergraduate Texts in Mathematics, Springer, New York, 2009, A computational approach. MR 2464052

[Sti03] John Stillwell, *Elements of number theory*, Undergraduate Texts in Mathematics, Springer-Verlag, New York, 2003. MR 1944957