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

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

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 $0, \ldots, b-1$ respectively. 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:

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

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 a few years ago, 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 third grade. 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. Working in base 3, compute $201 + 112$. (Give the answer in base 3, and state what all of these numbers are in decimal as well.)

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. 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 factor 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 $p_1 p_2 \cdots p_r$. 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 $i \neq j$ $(1 \leq i, j \leq k)$. 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 even, 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 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.[4] 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 $\mathbb{Z}$[5] as in Fig. 1.1.1. In terms of the picture for $\mathbb{N}$, 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.

[4] 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.

[5] $\mathbb{Z}$ denotes the integers, as you should have learned before, though I (quasi)formally introduce $\mathbb{Z}$ at the beginning of the next section.

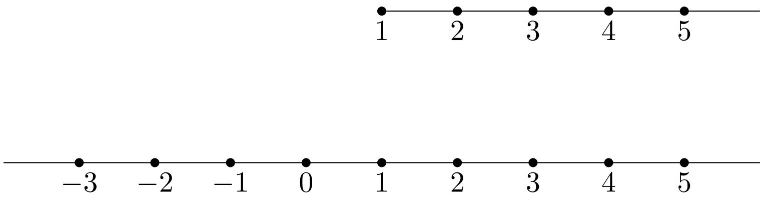


Figure 1.1.1: Visualizing $\mathbb{N}$ and $\mathbb{Z}$

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 $S_n \implies S_{n-1}$. Hence for any n we can reduce the problem 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.

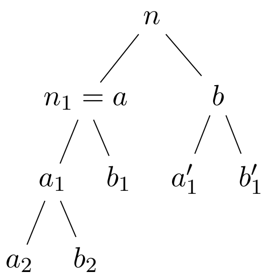
Of course in the above example, there is no real advantage to descent over induction. But it many situations, descent provides a way to think about a problem, which may be more intuitive than induction. Here is one that is of importance for us. (I'll leave it to you whether or not you think this is more 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 $p_1, \ldots, p_r$ 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 may be many ways one can break up n into factors, 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 divorces from representing a physical number of objects. This extends the whole numbers we know to give the **integers**:

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

(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.

Recall, 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 $\mathbb{Z} \times \mathbb{N}$.[6]
>
> (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., $\sqrt{2}$, as 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** $\mathbb{R}$.[7] (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 $\mathbb{Q}$, $\mathbb{R}$ also has 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**

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

[7] 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

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 $c_n z^n + c_{n-1} z^{n-1} + \cdots + c_1 z + c_0$ 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.

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 $x * y \in S$.*[8]

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

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

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.

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

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 $b \cdot \frac{a}{b} = a$ whenever $\frac{a}{b} \in \mathbb{Q}$.

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 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**[9] *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*:*

$$a(b+c) = ab + ac,^{10} \quad \textit{for all } a, b, c \in R; \ \textit{and} \tag{1.2.4}$$

(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[11] *and R satisfies the following property:*

(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.

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

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

[11] 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.

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*

(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.

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

Property (1) follows as

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

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.

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

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.[12] 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.

[12] 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.*[13] 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.,

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

[13] 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 ring and fields. One can of course check the definition by checking all the axioms.

Even if we want to be formal about proving things are rings or fields, in practice, we don't need to actually check all the axioms to show something is a ring or a field.

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 $S \subseteq R$ be non-empty. 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
>
> $$\mathbb{Z}[\frac{1}{2}] = \left\{ \frac{a}{b} \in \mathbb{Q} : b = 2^n \text{ for some } n \in \mathbb{Z}_{\geq 0} \right\}.$$
>
> 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 $\mathbb{Z}/n$.[14]

[14] Many 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 many number theorists don't like that notation. Actually, $\mathbb{Z}/n$

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 **??**. For now, we'll just explain the number systems $\mathbb{Z}/n\mathbb{Z}$.

For integers a, b, we write $a|b$ 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 $n|(b-a)$, 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*

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

(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 will typically 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:*

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

and

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

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.

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 $0 \geq a' < a$. 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., $n|(b - a)$. 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
>
> $$\begin{aligned} 2\mathbb{Z} &= \{\ldots, -4, -2, 0, 2, 4, 6, \ldots\} \\ 2\mathbb{Z} + 1 &= \{\ldots, -3, -1, 1, 3, 5, 7, \ldots\}. \end{aligned}$$

> **Example 1.3.2.** Suppose $n = 3$. Then the equivalence classes of $\mathbb{Z}$ mod 3 are
>
> $$\begin{aligned} 3\mathbb{Z} &= \{\ldots, -6, -3, 0, 3, 6, 9, \ldots\} \\ 3\mathbb{Z} + 1 &= \{\ldots, -5, -2, 1, 4, 7, 10, \ldots\} \\ 3\mathbb{Z} + 2 &= \{\ldots, -4, -1, 2, 5, 8, 11, \ldots\}. \end{aligned}$$

Definition 1.3.4. *Let $n \in \mathbb{Z}$. Denote the set of equivalence classes mod n by $\mathbb{Z}/n\mathbb{Z}$. 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 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.*

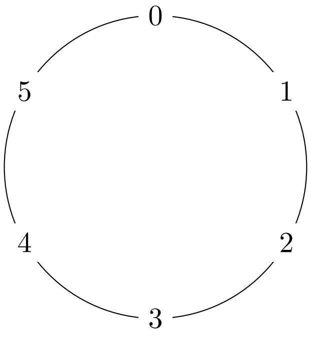


Figure 1.3.1: Visualizing $\mathbb{Z}/6\mathbb{Z}$

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.[15]

Proof. It is not to hard to check the ring axioms, e.g. ring axioms (1)–(3) follow in $\mathbb{Z}/n\mathbb{Z}$ because they do in $\mathbb{Z}$. 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 $a+b \equiv a'+b' \mod n$.

Since $a \equiv a' \mod n$ and $b \equiv b' \mod n$, 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 $a + b \equiv a' + b' \mod n$, 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. □

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

[15] 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 1.3.4. 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 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.

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 **??**.

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.,

$$10 + 11 + 12 + 13 + 14 \bmod 5,$$

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

$$10 + 11 + 12 + 13 + 14 \equiv 0 + 1 + 2 + 3 + 4 \equiv 1 + 2 + (-2) + (-1) \equiv 0 \bmod 5.$$

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

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

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.[16]

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.

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?

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

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.[17] 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

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	–	–	–	–	–

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

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,[18] 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, and thus 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.

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 $A = B = \text{hrair}$. Say $b = 5$. If we chose $a = 6$, we get $\text{hrair} - \text{hrair} = 1$, but if we chose $a = 7$ we get $\text{hrair} - \text{hrair} = 2$. 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 $\text{hrair} + \text{hrair} = \text{hrair}$ 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 we said that C is *the* equivalence class containing $a - b$, which implies that C is uniquely determined, but it is not always. (There is no issue with subtraction for modular arithmetic: think about it.)

[18] Rabscuttle or Blackberry might be.

> **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 only finitely many (but still larger than you can imagine) numbers, so that any numbers you would actually want to add or multiply you are still able 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[19], 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

[19] 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.

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 **??** (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 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 3.

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.[20] 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 4.

In this section, we will introduce quadratic rings[21] 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 nonsquare. 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. Why was the above so easy? Well, the above lemma is almost tautological, 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 nonsquare in $\mathbb{N}$ is a nonsquare in $\mathbb{Q}$. Hence if $d \in \mathbb{N}$ is a nonsquare, then $\sqrt{d}$ is irrational. More generally, if $d \in \mathbb{Z}$ is a nonsquare, 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 nonsquare (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

[20] 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 $i = \sqrt{-1}$.

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

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, $p|a$. 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 $4|a$ but $8 \nmid a$, then $16|d$ 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 $\mathbb{N}$. (*Suggestion:* 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 nonsquare. 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 nonsquare 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 nonsquare, $\mathbb{Z}[\sqrt{d}]$ 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 nonsquare in $\mathbb{Z}$, whence $d - (a/b)^2 \neq 0$ by the above lemma. □

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.

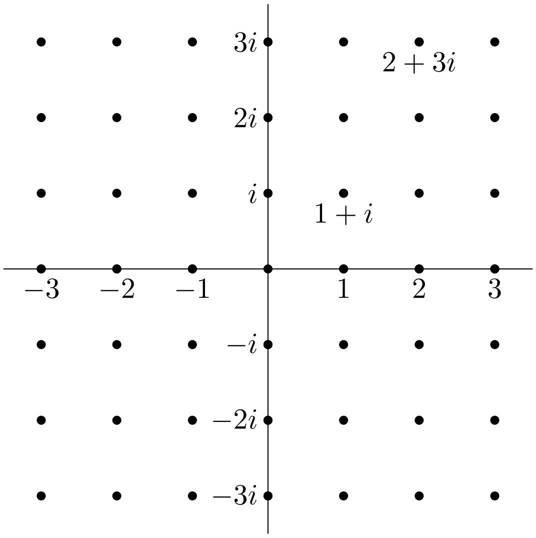


Figure 1.5.1: $\mathbb{Z}[i]$ inside $\mathbb{C}$

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 ring 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 $\mathbb{Z}[\sqrt{-3}] = \{a + \sqrt{3}bi : a, b \in \mathbb{Z}\}$ 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 $\sqrt{3}$.

Example 1.5.3. Let $d = -4$. Then $\mathbb{Z}[\sqrt{d}] = \mathbb{Z}[\sqrt{-4i}] = \mathbb{Z}[2i] = \{a + 2bi : a, b \in \mathbb{Z}\}$. 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 nonsquares. Show that $\mathbb{Z}[\sqrt{d'}]$ is a subring of $\mathbb{Z}[\sqrt{d}]$ if

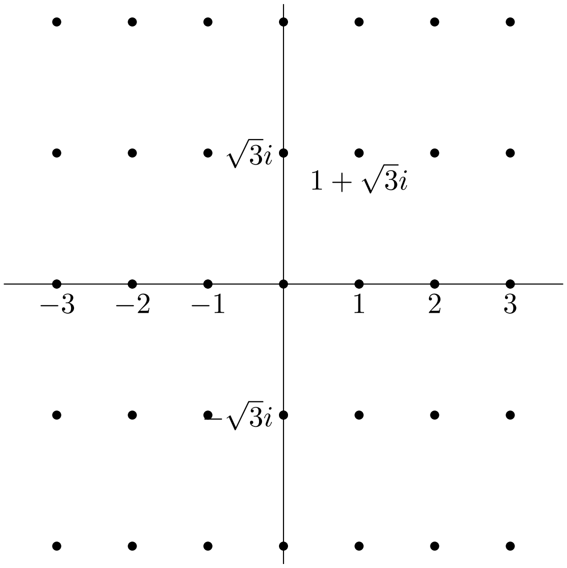


Figure 1.5.2: $\mathbb{Z}[\sqrt{-3}]$ inside $\mathbb{C}$

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 nonsquares. 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 nonsquares $d, d' \in \mathbb{Z}$ such that $\mathbb{Q}(\sqrt{d}) = \mathbb{Q}(\sqrt{d'})$ but neither $d|d'$ nor $d'|d$ is true.

Now let's look at a real quadratic example.

Example 1.5.4. Let $d = \sqrt{2}$. Then $\mathbb{Z}[\sqrt{2}] = \left\{a + b\sqrt{2} : a, b \in \mathbb{Z}\right\}$ 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.

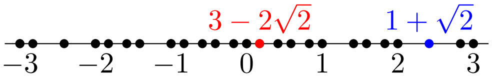


Figure 1.5.3: A sample of $\mathbb{Z}[\sqrt{2}]$ inside $\mathbb{R}$

Proposition 1.5.5. *Let $d \in \mathbb{Z}$ be a nonsquare.*

(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 nonsquare. 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

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

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 $N(a+b\sqrt{d}) = a^2 + |d|b^2 \leq n$ then necessarily $|a| \leq \sqrt{n}$ and $|b| \leq \sqrt{n}$ (in fact $|b| \leq \sqrt{\frac{n}{|d|}}$), 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 nonsquare. 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 $z = a + bi$ 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 $\sqrt{n}$, $1 \leq n \leq 8$ 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

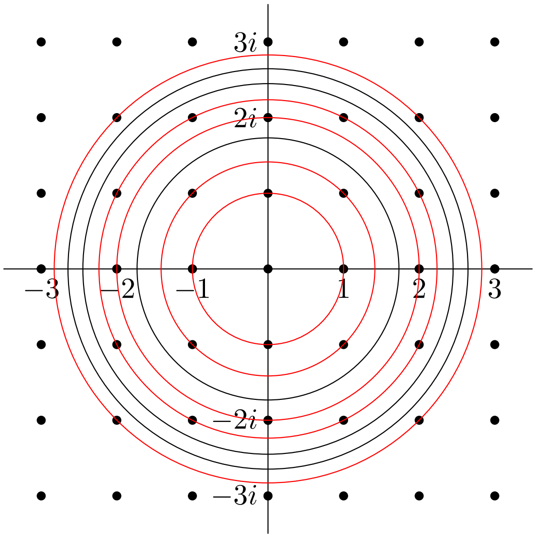


Figure 1.5.4: $\mathbb{Z}[i]$ with circles of radius $\sqrt{n}$, $1 \leq n \leq 8$

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: $\mu_1 = \{1\}$, $\mu_2 = \{1, -1\}$, $\mu_4 = \{1, -1, i, -i\}$.

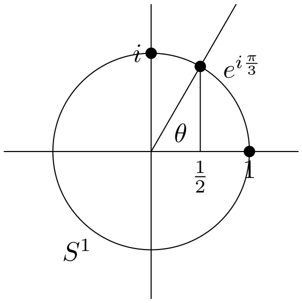
> **Exercise 1.6.1.** Show that if $m|n$, $\mu_m \subset \mu_n$.

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 $r \geq 0$. 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 $0 \leq \theta < 2\pi$ 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

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

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.)

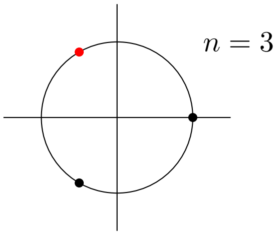
We will denote the first nontrivial solution on this list by:

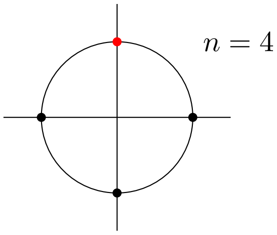
$$\zeta_n = e^{\frac{2\pi i}{n}}.$$

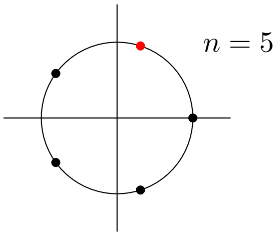
Then we can write the n-th roots of unity as

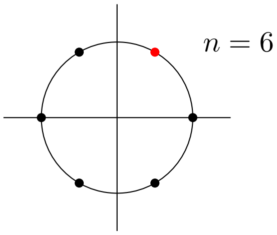
$$\mu_n = \left\{ e^{\frac{2\pi k i}{n}} : 0 \leq k < n \right\} = \left\{ 1, \zeta_n, \zeta_n^2, \ldots, \zeta_n^{n-1} \right\}.$$

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 $1, \zeta_n, \zeta_n^2, \ldots, \zeta_n^{n-1}$ 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 $1 + \zeta_2 = 0$.

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 were subrings of $\mathbb{C}$. You might wonder if there are any 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.

Another type of number 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

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

You can define addition, and extend the definition of multiplication to make $\mathbb{H}$ into a *non-commutative 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. Time permitting, we will explain this later in the course.

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, and they also have interesting application 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, this is such topics are not part of the present course.