Chapter 2

Factorization

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.1 Units, irreducibles and existence of factorizations

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The argument in the above example generalizes:

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

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

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

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

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

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

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

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

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

2.2 Primes and unique factorization

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

that are not the same up to ordering and units.

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

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

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

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

Then (2.2.1) implies

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

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

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

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

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

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

2.3 The Euclidean algorithm

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

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

The gcd by subtraction

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

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

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

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

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

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

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

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

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

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

Then we inductively compute

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

stopping only when we have

$$a_k = b_k.$$

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

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

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

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

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

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

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

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

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

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

The gcd by division with remainder

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

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

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

halting when we have a pair

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

Then

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

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

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

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

$$\gcd(16, 5) = \gcd(5, 3) = \gcd(3, 2) = \gcd(2, 1) = 1,$$

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

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

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

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

Linear Diophantine equations

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

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

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

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

$$ax + by = c,$$

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

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

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

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

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

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

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

□

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

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

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

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

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

We do this again to get

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

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

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

With one more step we are done:

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

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

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

which one already gets from the usual Euclidean algorithm, and

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

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

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

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

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

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

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

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

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

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

are precisely the set

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

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

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

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

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

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

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

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

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

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

Unique factorization for $\mathbb{Z}$

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

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

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

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

$$am + pn = 1.$$

Multiplying by b gives

$$abm + pbn = b.$$

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

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

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

2.4 A Euclidean algorithm for (two) quadratic rings

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

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

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

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

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

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

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

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

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

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

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

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

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

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

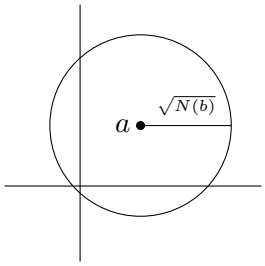
Definition 2.4.3. *Let $a, b \in \mathbb{Z}[\sqrt{d}]$, not both zero. We say d is a* **gcd** *(***greatest common divisor***) of a and b if $|N(d)|$ is maximal among elements such that $d|a$ and $d|b$. Denote the set of all gcds of a and b by* $\mathrm{GCD}(a, b)$.

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

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

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

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



Now the following norm inequality will be useful.

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

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

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

□

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

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

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

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

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

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

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

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

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

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

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

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

$$q = -1 - i, \quad r = 1, \quad N(r) = 1,$$
$$q = -1 - 2i, \quad r = i - 1, \quad N(r) = 2,$$
$$q = -i, \quad r = -2i, \quad N(r) = 4.$$

Note that not only are q, r are not unique, the norm of the remainders r are not even uniquely determined.

Exercise 2.4.1. Consider $a = 3 - \sqrt{-2}$, $b = 1 + \sqrt{-2}$ in $\mathbb{Z}[\sqrt{-2}]$. Find all $q, r \in \mathbb{Z}[\sqrt{-2}]$ such that $a = qb + r$ with $N(r) < N(b)$.

The next exercise shows $\mathbb{Z}[\sqrt{-3}]$ does not have the division property.

Exercise 2.4.2. Consider $a = 1 + \sqrt{-3}$, $b = 2$ in $\mathbb{Z}[\sqrt{-3}]$. Show there are no $q, r \in \mathbb{Z}[\sqrt{-3}]$ with $a = qb + r$ and $N(r) < N(b)$. (You can use the division algorithm but you don't have to.)

Theorem 2.4.5. *The rings $\mathbb{Z}[i]$ and $\mathbb{Z}[\sqrt{-2}]$ have the division property.*

Proof. Consider the ring $\mathbb{Z}[\sqrt{-d}]$ for $d > 0$. Let $a, b \in \mathbb{Z}[\sqrt{-d}]$ with $b \neq 0$. Any multiple qb of b in $\mathbb{Z}[\sqrt{-d}]$ is over the form $mb + ni\sqrt{d}b$ for some $m, n \in \mathbb{Z}$, i.e., the set of multiples qb of b are the lattice in $\mathbb{C}$ generated by b and $i\sqrt{d}b$. Recall multiplication by i acts as 90-degree rotation about the origin in $\mathbb{C}$, so $i\sqrt{d}b$ the point obtained rotating b 90-degrees and scaling by $\sqrt{d}$.

Now we use the lattice to tile $\mathbb{C}$ by rectangles whose vertices are lattice points qb, specifically translates of the rectangle (by mb and $ni\sqrt{d}b$ for $m, n \in \mathbb{Z}$) with vertices 0, b, $i\sqrt{d}b$ and $b + i\sqrt{d}b$. (See Fig. 2.4.1 for a picture when $d = 1$, $b = 2 + i$.) Each of these rectangles have side lengths $\sqrt{N(b)}$ and $\sqrt{dN(b)}$.

Now a lies in one of these rectangles R. Furthermore some vertex $v = qb$ of R lies within distance δ of a, where δ is one half of the diagonal length of R (the farthest an interior point of R can be from all vertices happens for the midpoint). It is easy to see $\delta = \frac{\sqrt{(1+d)N(b)}}{2}$. If $d \leq 2$, then $\delta < N(b)$. Consequently, $r = a - qb$ satisfies $N(r) < N(b)$. □

We wrote the argument above so you can see that it really only works for $d = 1, 2$. Moreover, if you think about the geometry of the argument, it seems like the division property should fail for $d \geq 3$. Indeed, we saw in Exercise 2.4.2 it fails for $d = 3$. However to prove this more generally, one needs to exhibit a rectangle R as in the above proof and an element of $\mathbb{Z}[\sqrt{-d}]$ (not just $\mathbb{C}$) which is farther away from every vertex of R than the shortest side length of R. I will simply leave the $d = 5$ case for you:

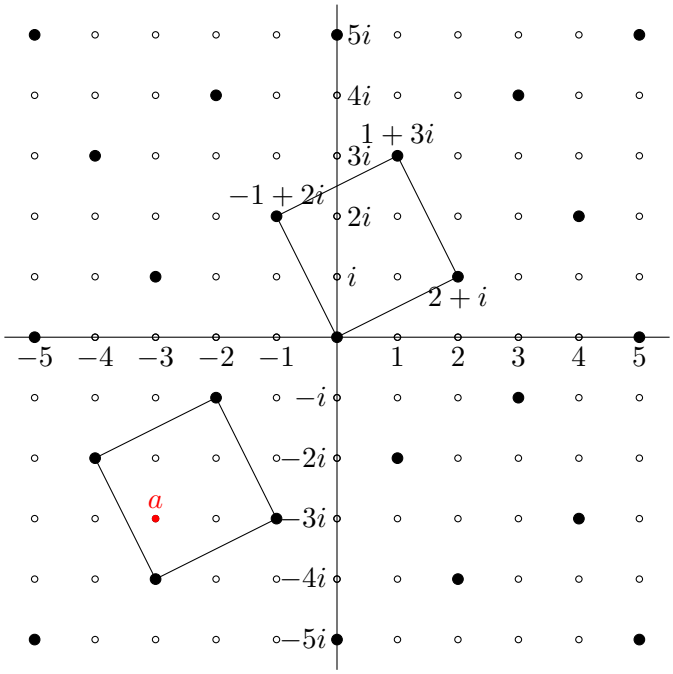


Figure 2.4.1: Multiples of $2+i \in \mathbb{Z}[i]$

Exercise 2.4.3. Show the division property does not hold in $\mathbb{Z}[\sqrt{-5}]$.

Now that we've established the division property for $\mathbb{Z}[i]$ and $\mathbb{Z}[\sqrt{-2}]$, we will see how the Euclidean algorithm extends to these cases.

Euclidean algorithm for $\mathbb{Z}[\sqrt{-d}]$ via division

Problem: Assume $\mathbb{Z}[\sqrt{-d}]$ satisfies the division property, e.g., $d = 1, 2$. Given a, b in $\mathbb{Z}[\sqrt{-d}]$ not both 0, find a gcd m of a and b.

(1) Let $\{a_1, b_1\} = \{a, b\}$ such that $N(a_1) \geq N(b_1)$. Put $i = 1$.

(2) If $b_i = 0$, we can take $m = a_i$ to be a gcd of a and b.

(3) Otherwise, apply the division algorithm to write $a_i = q_i b_i + r_i$ for some $q_i, r_i \in \mathbb{Z}[\sqrt{-d}]$ with $N(r_i) < N(b_i)$.

(4) Let $a_{i+1} = b_i$, $b_{i+1} = r_i$.

(5) Replace i with $i + 1$ and repeat from Step (2).

The proof that this works is similar to the case for $\mathbb{Z}$. With notation as in Step (3), note any divisor of both a_i and b_i is also a divisor of r_i, and conversly any divisor of r_i and b_i is a divisor of a_i, thus $\mathrm{GCD}(a_i, b_i) = \mathrm{GCD}(b_i, r_i) = \mathrm{GCD}(a_{i+1}, b_{i+1})$. Then by descent on

the norm of b_i, eventually some b_k is 0, whence we are led to $\mathrm{GCD}(a,b) = \mathrm{GCD}(a_k, 0)$ Now note that a_k is a gcd of a_k and 0 for any $a_k \neq 0$ so at some point we get a gcd $m = a_k$ in Step (2).

Alternatively, we can replace Step (2) by

(2′) If $b_i | a_i$, we can take $m = b_i$ to be a gcd of a and b.

This is true because we get $b_k = 0$ exactly when $b_{k-1} | a_{k-1}$. If $b_{k-1} | a_{k-1}$, both versions of the algorithm will output $a_k = b_{k-1}$ as a gcd, but the version using (2′) will just make one less pass. and it will simply terminate the algorithm earlier. However I originally wrote the algorithm with the formulation in Step (2) because (i) it is easier to formula descent terminating with $b_i = 0$ rather than the condition $b_i | a_i$, and (ii) to test for $b_i | a_i$ in general, you need to first apply the division algorithm in Step (3). That said, in some cases it will be obvious that $b_i | a_i$ (e.g., if $b_i = 1$ so 1 is a gcd) so when working out examples by hand we may use Step (2′).

Example 2.4.3. Consider $a = 2 - 3i$, $b = 1 + 2i$ in $\mathbb{Z}[i]$ as in Example 2.4.2. Since $N(a) = 13 > N(b) = 5$, we take $a_1 = a$, $b_1 = b$. By Example 2.4.2, there are 3 ways to write $a_1 = q_1 b_1 + r_1$. Let's see how the algorithm proceeds with these different choices.

If we take $q_1 = -1 - i$ then $r_1 = 1$, so we take $a_2 = b_1 = 1 + 2i$ and $b_2 = r_1 = 1$. Since $b_2 = 1$, Step (2') tells us 1 is a gcd of a and b.

Next, suppose instead we had taken $q_1 = -1 - 2i$ so $r_1 = i - 1$. Then $a_2 = 1 + 2i$ and $b_2 = i - 1$. Note $1 + 2i = (-i)(i - 1) + i$, so we can take $q_2 = -i$, $r_2 = i$, so $a_3 = i - 1$ and $b_3 = i$. Clearly $a_3 = 1 \cdot b_3 - 1$, so we can take $a_4 = i$ and $b_4 = -1$. Then Step (2') tells us -1 is a gcd of a and b. (In fact, for the second step we could've also taken $q_2 = 1 - i$ and $r_2 = 1$ which would give 1 as a gcd like in the previous case.)

Finally, suppose we had taken $q_1 = -i$ so $r_1 = -2i$. Then $a_2 = 1 + 2i$ and $b_2 = -2i$. We can write $1 + 2i = (-1)(-2i) + 1$, so we can take $q_2 = -1$, $r_2 = 1$, which gives $a_3 = -2i$ and $b_3 = 1$. Again this gives us a gcd of 1.

This example show that for given a, b the Euclidean algorithm can produce different sequences of a_i's and b_i's, as well result in as different elements of $\mathrm{GCD}(a,b)$, because the division algorithm provides multiple solutions to choose from at each stage. However, in the case that we have a Euclidean algorithm, or even just unique factorization, all gcd's of a and b differ by units, so this algorithm will always output the same final result up to a unit:

Exercise 2.4.4. Prove that if $\mathbb{Z}[\sqrt{-d}]$ has unique factorization, then for $a, b \in \mathbb{Z}[\sqrt{-d}]$ not both zero, $\mathrm{GCD}(a,b)$ is of the form $\{mu : u \text{ is a unit}\}$ for some fixed gcd m of a and b.

Exercise 2.4.5. Use the Euclidean algorithm to compute a gcd of $5 - 5i$ and $3 + 4i$ in $\mathbb{Z}[i]$.

2.5 Unique factorization beyond $\mathbb{Z}$

Now we come to the main results of this chapter beyond the case of $\mathbb{Z}$.

Theorem 2.5.1. *Let $d = 1, 2$. Then $\mathbb{Z}[\sqrt{-d}]$ satisfies the prime divisor property and has unique factorization.*

Proof. Recall from Theorem 2.2.5 that having the prime divisor property implies unique factorization (in fact by Exercise 2.2.3 they are equivalent), so it suffices to prove the prime divisor property. Now that we have a Euclidean algorithm, the proof is similar to that for $\mathbb{Z}$ in Theorem 2.3.3.

Let $a, b \in \mathbb{Z}[\sqrt{-d}]$ be nonzero nonunits. Suppose $p \in \mathbb{Z}[\sqrt{-d}]$ is irreducible and $p|ab$ but $p \nmid a$. To prove the prime divisor property, it suffices to show $p|b$. Since p is irreducible, its only divisors are of the form u and up where u is a unit of $\mathbb{Z}[\sqrt{-d}]$. Since $p \nmid a$, $up \nmid a$ for any unit u, hence $\mathrm{GCD}(p, a) = \{u : u \text{ is a unit}\}$.

Now, since we have a Euclidean algorithm for $d = 1, 2$, we can apply it to $\{a_1, b_1\} = \{p, a\}$. Then after one pass we have $a_2 = b_1$, $b_2 = r_1 = a_1 - q_1 b_1$ (in the notation of the algorithm in the previous section), i.e., either $\{a_2, b_2\} = \{a, p - q_1 a\}$ or $\{a_2, b_2\} = \{p, a - q_1 p\}$ (basically depending on whether $N(p) \geq N(a)$ or not). Inductively, it follows that at each step in this algorithm, a_i and b_i are $\mathbb{Z}[\sqrt{-d}]$-linear combinations of a and p, i.e., of the form $xa + yp$ for some $x, y \in \mathbb{Z}[\sqrt{-d}]$.

Since we eventually arrive at some a_k being a gcd $u \in \mathrm{GCD}(p, a)$, we have

$$ax + py = u,$$

for some $x, y \in \mathbb{Z}[\sqrt{-d}]$ and some unit u. (Replacing x, y with $u^{-1}x$, $u^{-1}y$, we could assume $u = 1$ if we want.) Multiplying by b gives

$$abx + pby = ub.$$

Since p divides the left hand side, $p|ub$, whence $p|b$. □

One of the main results of this course will be a proof of Fermat's classification of which numbers n are sums of two squares, i.e., when is $x^2 + y^2 = n$ solvable for $x, y \in \mathbb{Z}$? We will use unique factorization in $\mathbb{Z}[i]$ to prove this, and similarly one can use unique factorization in $\mathbb{Z}[\sqrt{-2}]$ to determine when $x^2 + 2y^2 = n$ has a solution. We will get to this in Chapter 4 after first treating modular arithmetic in Chapter 3.

The reason for this ordering is that I'm trying to roughly organize the chapters by topics, and wanted to prove the fundamental theorem of arithmetic before doing modular arithmetic, as we will talk about factorization problems related to cryptography in that chapter. But we will also use modular arithmetic in the determination of numbers which are sums of 2 squares, so that needed to go before we get our main application of Theorem 2.5.1. However, since we worked so long to get this theorem, I'd like to show you know how to get a nice application right now. Here is one, though I leave some details to the exercises.

Definition 2.5.2. *Let $d \in \mathbb{Z}$ and $a, b \in \mathbb{Z}[\sqrt{d}]$. We say a and b are* **relatively prime** *(or* **coprime***) if 1 is a gcd of a and b.*

Proposition 2.5.3. *The only solution to $y^3 = x^2 + 2$ in $\mathbb{N}$ is $(x, y) = (5, 3)$.*

Proof. Suppose

$$y^3 = x^2 + 2 = (x + \sqrt{-2})(x - \sqrt{-2}).$$

It is not hard to show that $x + \sqrt{-2}$ and $x - \sqrt{-2}$ are relatively prime in $\mathbb{Z}[\sqrt{-2}]$ (first exercise below). Since their product is a cube (i.e., of the form a^3 for some $a \in \mathbb{Z}[\sqrt{-2}]$), then both $x + \sqrt{-2}$ and $x - \sqrt{-2}$ are cubes by unique factorization in $\mathbb{Z}[\sqrt{-2}]$ (second exercise below). Write

$$x + \sqrt{-2} = (m + n\sqrt{-2})^3 = m^2 - 6mn^2 + (3m^2n - 2n^3)\sqrt{-2}$$

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

$$x = m^3 - 6mn^2\, 1 = n(3m^2 - 2n^2).$$

From the second equation, we have $n = \pm 1$ and $3m^2 - 2n^2 = 3m^2 - 2 = 1$, so $m = \pm 1$ and $x = \pm 5$. □

Exercise 2.5.1. Suppose $x, y \in \mathbb{N}$ such that $y^3 = x^2 + 2$. This is how to prove $x + \sqrt{-2}$ and $x - \sqrt{-2}$ are coprime in $\mathbb{Z}[\sqrt{-2}]$.
(i) Show x must be odd.
(ii) Show for any odd $x \in \mathbb{Z}$, $x+\sqrt{-2}$ and $x-\sqrt{-2}$ are coprime in $\mathbb{Z}[\sqrt{-2}]$. (*Suggestion:* Try adding and subtracting these two quantities.)

Exercise 2.5.2. Let $d = 1, 2$. Use Theorem 2.5.1 to show that if ab is a cube in $\mathbb{Z}[\sqrt{-d}]$ and a and b are relatively prime, then a and b are cubes in $\mathbb{Z}[\sqrt{-d}]$.

Now that we've seen some concrete utility of unique factorization in quadratic rings, you might wonder for what other quadratic rings one gets unique factorization. You might be worried from our proof of the division algorithm that we don't actually get unique factorization for any other imaginary quadratic rings. However, that's not exactly true.

Let's recall how various properties of quadratic rings are related:

$$\begin{aligned} &\text{division property} \implies \\ &\text{Euclidean algorithm} \implies \\ &\text{prime divisor property} \iff \\ &\text{unique factorization} \end{aligned}$$

So we while we suggested that $\mathbb{Z}[\sqrt{-d}]$ does not have the division property for $d \geq 3$, that doesn't mean we need that or the Euclidean algorithm to have prime factorization, but only that it's a useful tool for proving unique factorization when $d = 1, 2$.

Let's think about the next example, $d = 3$. We saw in Exercise 2.4.2 that $\mathbb{Z}[\sqrt{-3}]$ does not have the division property, and we also saw in Exercise 2.2.2 that $\mathbb{Z}[\sqrt{-3}]$ does not have unique factorization, namely

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

are two distinct factorizations of 4 into irreducibles of $\mathbb{Z}[\sqrt{-3}]$. However, in some sense, the problem with this example is simply that $\mathbb{Z}[\sqrt{-3}]$ does not have enough elements, and

this can be resolved by passing to the Eisenstein integers $\mathbb{Z}[\zeta_3] = \mathbb{Z}[\frac{1+\sqrt{-3}}{2}]$, which contains $\mathbb{Z}[\sqrt{-3}]$. Namely, in $\mathbb{Z}[\zeta_3]$ we can factor

$$2 = \frac{1+\sqrt{-3}}{2} \cdot (1 - \sqrt{-3}),$$

so the above factorizations of 4 in $\mathbb{Z}[\sqrt{-3}]$ resolve into the factorizations

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

Indeed one has unique factorization in $\mathbb{Z}[\zeta_3]$, and one can prove it in a similar manner to above. The key step is the division property:

Exercise 2.5.3. Show $\mathbb{Z}[\zeta_3]$ has the division property. (*Suggestion:* First think about the rectangles we constructed for $\mathbb{Z}[\sqrt{-3}]$ in the proof of Theorem 2.4.5. Then think about what multiples you get inside these rectangles if you can multiply by $\frac{1+\sqrt{-3}}{2}$.)

Theorem 2.5.4. *The Eisenstein integers $\mathbb{Z}[\zeta_3]$ have unique factorization.*

Proof. Once one has the division property, one gets a Euclidean algorithm, and the proof is the same as that for $\mathbb{Z}[i]$ and $\mathbb{Z}[\sqrt{-2}]$. □

Unique factorization for $\mathbb{Z}[\zeta_3]$ is useful for proving Fermat's Last Theorem for $n = 3$, i.e., $x^3 + y^3 = z^3$ has no nontrivial solutions (i.e., no solutions with x, y, z all nonzero). We'll discuss this in Chapter 6.

Exercise 2.5.4. Even though $\mathbb{Z}[\zeta_3]$ has unique factorization, show that there exist $a, b \in \mathbb{Z}[\zeta_3]$ which are relatively prime with ab a cube but neither a nor b are cubes. What is the difference between this situation and Exercise 2.5.2?

Now that we've seen $\mathbb{Z}[\zeta_3]$ can resolve failure of unique factorization in $\mathbb{Z}[\sqrt{-3}]$, you might wonder if you can do this for other rings, e.g., $\mathbb{Z}[\sqrt{-5}]$ or $\mathbb{Z}[\sqrt{-7}]$? It turns out you can in one but not the other. For $\mathbb{Z}[\sqrt{-3}]$ this worked because we could adjoin "quadratic integer" $\frac{1+\sqrt{-3}}{2}$, and then the geometry works out nice in Exercise 2.5.3.

Regarding the quadratic integer terminology, what it means in general for a complex number z to be an **algebraic integer** is that it is a root of some monic polynomial $z^n + c_{n-1}z^{n-1} + \cdots c_1 z + c_0$ for some $c_i \in \mathbb{Z}$. E.g., since ζ_n^j satisfies $z^n - 1 = 0$, each root of unity ζ_n^j is an algebraic integer. So are sums, differences and products of algebraic integers (the set of all algebraic integers forms a ring). In particular $\zeta_3 = \frac{-1+\sqrt{-3}}{2}$ is an algebraic integer, and so is $1 + \zeta_3 = \frac{1+\sqrt{-3}}{2}$.

We say $d \in \mathbb{Z}$ is **squarefree** if $m \in \mathbb{N}$ with $m^2 | d$ implies $m = 1$, i.e., d is not nontrivially divisible by any squares. Note that if $d \in \mathbb{Z}$ with $d = d_0 m^2$ where d_0 is squarefree and $m \in \mathbb{N}$, then $\sqrt{d} = m\sqrt{d_0}$ so

$$\mathbb{Z}[\sqrt{d}] = \left\{a + bm\sqrt{d_0} : a, b \in \mathbb{Z}\right\} = \mathbb{Z}[m\sqrt{d_0}]$$
$$\mathbb{Q}(\sqrt{d}) = \left\{a + bm\sqrt{d_0} : a, b \in \mathbb{Q}\right\} = \mathbb{Q}(\sqrt{d_0}).$$

So when we want to work with quadratic rings which are as big as possible (and thus increasing the likelihood of unique factorization, admittedly for reasons I haven't completely explained), we may as well restrict to d squarefree. Moreover, one gets all quadratic fields $\mathbb{Q}(\sqrt{d})$ by restricting to squarefree d.

Definition 2.5.5. *Let $d \in \mathbb{Z}$ be squarefree, $d \neq 0, 1$. The* **ring of integers** *of $\mathbb{Q}(\sqrt{d})$ is*

$$\mathcal{O}_d = \begin{cases} \mathbb{Z}[\frac{1+\sqrt{d}}{2}] = \left\{a + b \cdot \frac{1+\sqrt{d}}{2} : a, b \in \mathbb{Z}\right\} & \text{if } d \equiv 1 \bmod 4, \\ \mathbb{Z}[\sqrt{d}] & \text{else.} \end{cases}$$

Exercise 2.5.5. Show each $\mathcal{O}_d$ is a ring.

There is a general result that says that (for d any nonsquare) $\mathbb{Z}[\sqrt{d}]$ can only have a Euclidean algorithm or unique factorization if $\mathbb{Z}[\sqrt{d}]$ is a full ring of integers $\mathcal{O}_d$. Necessarily, d is both squarefree and $d \not\equiv 1 \bmod 4$.

We now give a brief summary of what is known. Gauss, in his *Disquistiones Arithmeticae* from 1801 (when he was only 23), which was a major milestone in number theory, found that imaginary quadratic rings of integers $\mathcal{O}_d$ ($d < 0$) have unique factorization when

$$d = -1, -2, -3, -7, -11, -19, -43, -67, -163$$

and conjectured there are no others. On the other hand, Gauss also conjectured that there are infinitely many real quadratic rings of integers $\mathcal{O}_d$ ($d > 0$) with unique factorizations.

Gauss's conjecture on imaginary quadratic rings was proved in the 1960's by Heegner (a high school teacher) and Stark, and independently by Baker. Gauss's conjecture on real quadratic rings is still one of the major open problems in number theory, despite that unique factorization seems extremely common for real quadratic $\mathcal{O}_d$. (More precisely, Cohen and Lenstra conjectured $\mathcal{O}_d$ should have unique factorization for more than 75% of squarefree $d > 1$.)

In any case, we see that having unique factorization is a rather special property for rings we care about in number theory—e.g., it only happens for finitely many imaginary quadratic rings. We also remark that it only happens finitely many cyclotomic rings $\mathbb{Z}[\zeta_n]$ (the largest of which is with $n = 90$). Much of algebraic number theory was developed as an attempt to understand unique factorization and how to work in number rings when unique factorization fails. There are three main ways to "resolve" nonunique factorization: (i) Gauss's theory of binary quadratic forms for quadratic rings; (ii) Kummer's theory of ideal numbers (basically, work in a larger ring S than your given ring R, where unique factorization still might not hold, but that any factorization of R into irreducibles of S is essentially unique); (iii) Dedekind's ideal theory (a generalization of Kummer's ideal numbers, where one work with certain sets instead of number). Of these (iii) is by far the most common theory to use regarding factorization in number rings in modern number theory. Now we're not studying any of these in this class—(ii) and (iii) are more advanced than what we will do in this course, whereas (i) goes in a different direction.

Going back to our musings on which imaginary quadratic rings have a Euclidean algorithm or unique factorization, the proof of Gauss's conjecture in the imaginary case tells

us that $\mathbb{Z}[\sqrt{-d}]$ ($d > 0$) only has unique factorization when $d = 1, 2$, the cases we proved. To see this, note the other cases where $\mathcal{O}_{-d}$ has unique factorization with $d > 0$ fall in the situtation $-d \equiv 1 \bmod 4$, i.e., $d \equiv 3 \bmod 4$, so $\mathbb{Z}[\sqrt{-d}] \neq \mathcal{O}_{-d}$, and use the "general result" mentioned above. Consequently, the only cases where $\mathbb{Z}[\sqrt{-d}]$ can have a Euclidean algorithm is when $d = 1, 2$, and we proved in these cases it does.

One does not see this by just looking at the rings $\mathbb{Z}[\sqrt{-d}]$, but having a Euclidean algorithm really is more special than just having unique factorization. For the imaginary quadratic rings of integers $\mathcal{O}_{-d}$, they only have Euclidean algorithms if $d = -1, -2, -3, -7$, or -11. Thus the last 4 cases on Gauss's imaginary quadratic list have unique factorization but no Euclidean algorithm. Among real quadratic rings $\mathcal{O}_d$, we only have a Euclidean algorithm (with respect to the norm[4]) when

$$d = 2, 3, 5, 6, 7, 11, 13, 17, 19, 21, 29, 33, 37, 41, 57, 73.$$

[4]There is a more general notion of a ring having a "Euclidean algorithm," namely one can measure size not just using the norm, but possibly using another function. It is conjectured that infinitely many real quadratic rings $\mathcal{O}_d$ have a Euclidean algorithm in this more general sense, but still there are $\mathcal{O}_d$ which have unique factorization but do not have a generalized Euclidean algorithm.