Unique Factorization Domains

Iurie Boreico

Extended notes from number theory lectures at AwesomeMath Camp 2008

1 Introduction

The key concept in number theory is the concept of divisibility. With the help of factorization, the tools of divisibility are fundamental in attacking the vast majority of the problems in elementary number theory. However, as the difficulty level of the problems increases, it may become impossible to factorize an expression, at least over the integers. However, it might be possible to factorize it over some other, larger, set, like $\mathbb{R}$ or $\mathbb{C}$. If that's the case, one might try to adapt the already familiar principles of divisibility to this new set, for sake of using their power to the problem. Now, it turns our that $\mathbb{R}$ and $\mathbb{C}$ are sets that are too large and too far from the usual integers, and therefore an adequate theory of divisibility in them can not be developed (or at least, it will be so different from the integer divisibility that it will be of no help). This is why one should consider smaller sets, which bear enough resemblance to $\mathbb{Z}$ so that a theory of divisibility can be constructed. In this note, we will analyze the simplest possible extensions of $\mathbb{Z}$, (quadratic fields) and study the divisibility in these sets. The results obtained will be helpful for solving a class of diophantine equations and establishing some beautiful theorems of number theory. But more importantly, these quadratic fields may be regarded as a window into algebraic number theory, an extremely rich branch of mathematics that has arisen exactly from the need to construct divisibility laws in algebraic extensions on $\mathbb{Q}$, in order to solve some diophantine equations (most importantly, Fermat's Last Theorem).

Remark. We do not want to go very deep here and wish to minimize the level of abstraction. This is why some notions from algebra are simplified, for example rings are assumed to be commutative, and whatever advanced concepts are introduced, they are presented to mirror some already familiar concepts from elementary number theory.

2 Rings, fields, and other words with multiple meanings

The set of integers, $\mathbb{Z}$, beside being a set, has some considerable properties. We can add, subtract, multiply elements from $\mathbb{Z}$ and these operations will produce as result elements from $\mathbb{Z}$. It is thus natural to expect, that whatever we want to extend $\mathbb{Z}$ to, this bigger set also have this characteristics. They are encoded in the definition of a ring.

Definition 2.1. A ring $R(+,\cdot)$ consists of a set R together with two operations: $+$ (addition) and $\cdot$ (multiplication) which have the properties:

i) For all $a, b \in R$, $a+b$ and $a \cdot b$ are elements of R.

ii) The operation $+$ has an inverse $-$, that is, for any $a, b \in R$, there exists a unique element $c = a - b$ such that $c + b = a$.

iii) The operations $+$ and $\cdot$ are associative, that is, for any $a, b, c \in R$, one has $(a+b)+c = a+(b+c)$ and $(a \cdot b) \cdot c = a \cdot (b \cdot c)$. We denote these numbers by $a+b+c$, respectively abc.

iv) The operations $+, \cdot$ are commutative, that is $a+b = b+a$ and $ab = ba$ for all $a, b \in R$.

v) The addition is distributive with respect to multiplication, that is $a(b+c) = ab + ac$ and $(b+c)a = ba + ca$. In other words, we can just open the brackets in expressions just like we are used to.

vi) Both addition and multiplication have neutral elements, that is, there exist elements denoted by 0 and 1 such that $x + 0 = 0 + x = x$ for all $x \in R$ and $x \cdot 1 = 1 \cdot x = x$ again for all $x \in R$.

This looks like a long string of properties, but it should not look very intimidating. These are just the basic properties of addition and multiplication that we all know by heart from our math experience. In fact, the very notion of ring was inspired by the example of $\mathbb{Z}$, and $\mathbb{Z}$ was the first known ring.

Let's see some examples of rings. Beside $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ are also rings, since the ability to add and multiply in these sets is already carved deep in our souls. Also $\mathbb{Z}/n\mathbb{Z}$, the set of residues modulo n, and $\mathbb{Z}[X]$, the set of polynomials with integer coefficients, are rings too. For that matter, $R[X]$, the set of all polynomials with coefficients in a ring R, is also a ring, and I leave to you as an exercise to prove it.

$\mathbb{Z}$ has a downside: one cannot always divide in it. And this is the problem of rings in general. But in some rings, like $\mathbb{R}$ or $\mathbb{Q}$, one can always perform division. These rings are called fields.

Definition 2.2. A field F is a ring in which the multiplication is invertible, that is, for every $x \in F, x \neq 0$, there exists an element y such that $xy = yx = 1$.

This element is denoted by x^{-1}, or $\frac{1}{x}$, and generally $\frac{a}{b} = ab^{-1}$ is the unique element c such that $bc = a$.

From our previous examples $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ are rings, while $\mathbb{Z}$ and $\mathbb{Z}[X]$ are not. If n is a prime, then $F_n = \mathbb{Z}/n\mathbb{Z}$, is also a field, since we know every non-zero residue has a inverse modulo p. If, however, n is not a prime, then this is not true: taking $n = ab$ where $a, b > 1$, the residue $\hat{a}$ has no inverse modulo n: if $\hat{a}\hat{c} = \hat{1}$ then $ac = nk + 1 = abk + 1$ for some integer k, which is impossible since the left-hand side is divisible by a while the right-hand side is clearly not.

The mnemonic rule for distinguishing rings and fields is simple: in fields we can divide, in rings - not. It useful to always imagine $\mathbb{Z}$ when thinking of a ring and $\mathbb{Q}$ when thinking of a field. Of course, this way of thinking is generally flawed, since $\mathbb{Z}$ is very different from $\mathbb{Z}/n\mathbb{Z}$, primarily because the latter is finite and the former is not. But for now, we will deal only with rings that are very similar to $\mathbb{Z}$, so you can compare every ring to it and be safe for the moment.

3 Quadratic Rings

The world of rings and fields is immeasurably rich, and the multitude of examples already shows that. However we deal here only with a special class of rings, called quadratic.

Consider, for example, the equation $x^2 - 2y^2 = 7z^2$. We cannot factor the left-hand side, which is a serious obstacle for solving this equation. And the reason we cannot do this is because $x^2 - 2y^2 = (x - \sqrt{2}y)(x + \sqrt{2}y)$, but $\sqrt{2}$ is not an integer. If we could add $\sqrt{2}$ to $\mathbb{Z}$, we would be able to factor it, and this is how quadratic rings are constructed.

Given a ring R and an "object" x which is not in R, one can try to construct a new, larger, ring R_1 by adding x to the ring. Now, since R_1 must be a ring, $x \in R_1$ implies immediately that $x^2 = x \cdot x$ is also in R_1, and by induction $x^k \in R_1$ for all $k \in \mathbb{N}$. And hence for any $a_k \in R$, $a_k x^k \in R_1$, so any sum of the form $\sum_{k=0}^{m} a_k x^k$ must be in R_1, which means R_1 must contain all expressions of form $\sum_{k=0}^{m} a_k x^k$.

As it is easy to prove, the set of all expressions of form $\sum_{k=0}^{m} a_k x^k$ is in fact, a ring, and from what we saw it is the smallest ring containing R and x. This process of enlarging the ring by adding a new element x is called "adjoining x".

Definition 3.1. Given a ring R, and an object $x \notin R$, one defines $R[x]$ to be the smallest ring containing R and x. $R[x]$ is the set of all expressions of form $\sum_{k=0}^{m} a_k x^k$. One says that x generates $R[x]$ over R, or simply that x is a generator (of $R[x]$ over R).

The concept of object is rather vague. We could have x any variable, in which case $R[x]$ will simply be the set of polynomials in x with coefficients in R. However one may impose some restrictions on x, say $x^2 = a \in R$, or any such polynomial

condition on x. In this case $R[x]$ will be much thinner: for example if $x^2 = a$, then $x^{2k} = a^k, x^{2k+1} = a^k x$, so any power of x can be transformed into an element of R or x times an element of R, which means $R[x]$ consists of just the expressions $ax+b$, $a, b \in R$.

In the above example, one should add $\sqrt{2}$ to $\mathbb{Z}$. One will obtain the ring $\mathbb{Z}[\sqrt{2}]$, consisting of expressions of form $a + b\sqrt{2}$, and in this ring, the expression $x^2 - 2y^2$ will factor into $(x - \sqrt{2}y)(x + \sqrt{2}y)$.

One should check that $\mathbb{Z}[\sqrt{2}] = \{a + b\sqrt{2} \mid a, b \in \mathbb{Z}[\sqrt{2}]\}$, is a ring. Indeed, we have $(a+b\sqrt{2})+(c+d\sqrt{2}) = (a+c)+(b+d)\sqrt{2} \in \mathbb{Z}[\sqrt{2}]$ and $(a+b\sqrt{2})(c+d\sqrt{2}) = ac + 2bd + (bc + ad)\sqrt{2} \in \mathbb{Z}[\sqrt{2}]$, so $\mathbb{Z}[\sqrt{2}]$ is stable under both multiplication and addition. The other ring properties are quite standard to verify.

Definition 3.2. Given an element ϵ which is a root of a monic quadratic x^2+ax+b in $\mathbb{Z}[X]$, the ring $\mathbb{Z}[\epsilon]$ is called a quadratic ring. It consists of expressions of form $x + y\epsilon$ where $x, y \in \mathbb{R}$.

The condition that ϵ is a root of x^2+ax+b is crucial for the set $\{x+y\epsilon \mid x, y \in \mathbb{Z}\}$ to be a ring. Indeed, $\epsilon^2 = -a\epsilon - b$, hence

$$\begin{aligned}(x + y\epsilon)(m + n\epsilon) &= xm + (ym + nx)\epsilon + yn\epsilon^2 \\ &= mn + (ym + nx)\epsilon - yn(a\epsilon + b) \\ &= mn - ynb + (ym + nx - yna)\epsilon\end{aligned}$$

and the last expression belongs to $\mathbb{Z}[\epsilon]$. Without the condition that a, b be integers, we would have $\epsilon^2 = -a\epsilon - b$ not belonging to $\mathbb{Z}[\epsilon]$.

Let us simplify $\mathbb{Z}[\epsilon]$ a bit. We know that $\epsilon = \frac{-a\pm\sqrt{a^2-4b}}{2}$. If we denote $D = a^2-4b$, then D must not be a perfect square (otherwise ϵ would belong to $\mathbb{Z}$). We claim $\mathbb{Z}[\epsilon] = \mathbb{Z}[\epsilon - k]$ for any integer k. Indeed, $m+n\epsilon = (m+nk)+n(\epsilon-k))$, and $\epsilon-k$ is a root of the equation $(x+k)^2+a(x+k)+b = 0$. Now $\epsilon - k = \frac{a-2k+\sqrt{D}}{2}$. So we can eliminate a completely, by setting $k = [\frac{a}{2}]$. Thus $\mathbb{Z}[\epsilon] = \mathbb{Z}[\frac{\sqrt{D}}{2}]$ or $\mathbb{Z}[\epsilon] = \mathbb{Z}[\frac{1+\sqrt{D}}{2}]$, depending whether a is even or odd.

If a is even, then $D = a^2 - 4b$ is divisible by 4 so by setting $d = \frac{D}{4}$ we get $\mathbb{Z}[\epsilon] = \mathbb{Z}[\sqrt{d}\,]$.

If a is odd, then $D = a^2 - 4b$ is congruent to 1 modulo 4. Setting $d = D$ we get $\mathbb{Z}[\epsilon] = \mathbb{Z}[\frac{1+\sqrt{d}}{2}]$. We thus have established the following theorem:

Theorem 3.3. All quadratic rings are of form $\mathbb{Z}[\sqrt{d}]$ for d not a perfect square or $\mathbb{Z}[\frac{1+\sqrt{d}}{2}]$ for $d \equiv 1 \pmod 4$ not a perfect square.

It is an exercise for you to check directly that these are indeed rings.

Note that these rings are all distinct. If $d_1 \neq d_2$ and $|d_1| \leq |d_2|$ then $\sqrt{d_1} \notin \mathbb{Z}[\sqrt{d_2}]$. Indeed, assume that $\sqrt{d_1} = a + b\sqrt{d_2}$. Then $d_1 = a^2 + b^2 d_2 + 2ab\sqrt{d_2}$ so

$2ab = 0$ and $b = 0$ or $a = 0$. If $b = 0$ then $d_1 = a^2$ contradicting the assumption that d_1 is not a perfect square. If $a = 0$ then $d_1 = b^2 d_2$ and $b^2 > 1$ hence $|d_1| > |d_2|$, contradiction. A similar but messier argument deals with rings of form $\mathbb{Z}[\frac{1+\sqrt{d}}{2}]$, but we leave it as an exercise.

Notice that every number d can be written uniquely as $d'k^2$ where d' is square-free. Then $\mathbb{Z}[\sqrt{d}] = \mathbb{Z}[k\sqrt{d'}]$ is a subset of the ring $\mathbb{Z}[\sqrt{d'}]$. Similarly, $\mathbb{Z}[\frac{1+\sqrt{d}}{2}]$ is a subset of the ring $\mathbb{Z}[\frac{1+\sqrt{d'}}{2}]$. Finally, if $d \equiv 1 \pmod 4$ then $\mathbb{Z}[\sqrt{d}]$ is a subset of $\mathbb{Z}[\frac{1+\sqrt{d}}{2}]$. So it suffices to consider only the cases when d is square-free. Otherwise we can enlarge the ring and why work with a smaller ring if we can deal with a larger one? Similarly, if $d \equiv 1 \pmod 4$, it is better to consider $\mathbb{Z}[\frac{1+\sqrt{d}}{2}]$ instead of the smaller $\mathbb{Z}[\sqrt{d}]$. Therefore it suffices to regard only the rings $R = \mathbb{Z}[\sqrt{d}]$ for square-free $d \equiv 2, 3 \pmod 4$ and $R = \mathbb{Z}[\frac{1+\sqrt{d}}{2}]$ for square-free $d \equiv 1 \pmod 4$. (Remember that d divisible by 4 is not square-free).

Remark. It is very important to understand the case $d \equiv 1 \pmod 4$ which gives rise to the somewhat weirder rings $\mathbb{Z}[\frac{1+\sqrt{d}}{2}]$. Indeed, if not for this case, all quadratic rings would have the nice and comfy form $\mathbb{Z}[\sqrt{d}]$. The cases $d \equiv 1 \pmod 4$ are uglier and may result in nastier computations. Nevertheless, they should be understood and accepted too, as they are not of lesser importance than the others.

If ϵ is a root of a quadratic equation $x^2 + ax + b$, then $\overline{\epsilon} = -a - \epsilon = \frac{b}{\epsilon}$ is also a root of the same quadratic equation, and it belongs to R, too. The elements ϵ and $\overline{\epsilon}$ are called conjugate. In some sense, ϵ and $\overline{\epsilon}$ are indistinguishable, because they satisfy the same condition with respect to $\mathbb{Z}$. We only know about ϵ that it is a root of $x^2 + ax + b$, and this condition does not differentiate ϵ from $\overline{\epsilon}$. The fact that they are so closely related is very important, since conjugation plays a very important role.

If $\epsilon = a + b\sqrt{d}$ $(a, b \in \mathbb{Q})$ then $\overline{\epsilon} = a - b\sqrt{d}$. From here, it is straightforward to check that $\overline{x} + \overline{y} = \overline{x + y}$. Similarly, if $x = a + b\sqrt{d}, y = u + v\sqrt{d}$, then

$$\begin{aligned}
\overline{x} \cdot \overline{y} &= (a - b\sqrt{d})(u - v\sqrt{d}) \\
&= au + bvd - (bu + av)\sqrt{d} \\
&= \overline{au + bvd + (bu + av)\sqrt{d}} \\
&= \overline{(a + b\sqrt{d})(u + v\sqrt{d})} \\
&= \overline{x \cdot y}.
\end{aligned}$$

Finally, $\overline{(\overline{x})} = x$ and note that $x = \overline{x}$ if and only if $x \in \mathbb{Z}$.

As we will see, the concept of conjugate is very useful. So is the concept of norm.

Definition 3.4. For $x \in R$, the norm of x is defined to be $N(x) = |x\overline{x}|$. For $x \notin \mathbb{Z}$, it is the absolute value of the free term of the quadratic equation satisfied by x. If $x \in \mathbb{Z}$, then $N(x) = x^2$.

The absolute value is applied only to ensure that the norm is positive, so we can use it as some sort of absolute value. Note that $N(x) > 0$ unless $x = 0$: if $N(x) = 0$ then $x\overline{x} = 0$ so either $x = 0$ or $\overline{x} = 0$, and if $\overline{x} = 0$ by conjugating we get $x = \overline{0} = 0$.

The norm has some nice properties:

Theorem 3.5. For a quadratic ring R, the norm of any number is an integer, and is a multiplicative function, that is, $N(x)N(y) = N(xy)$. Also $N(x) = N(\overline{x})$.

Proof. From the definition, $N(a + b\sqrt{d}) = |a^2 - db^2|$. It is clearly an integer in the case $d \equiv 2, 3 \pmod 4$, since a, b are integers. In the case $d \equiv 1 \pmod 4$, a and b may not be integers. If they aren't, then $2a, 2b$ are odd integers hence $a^2 - db^2 = \frac{(2a)^2 - d(2b)^2}{4}$ and $(2a)^2 - d(2b)^2$ is divisible by 4 so the norm is an integer. Let $x = a + b\sqrt{d}, y = u + v\sqrt{d}$ then $xy = (a + b\sqrt{d})(u + v\sqrt{d}) = (au + dbv) + (bu + av)\sqrt{d}$. Thus

$$\begin{aligned} N(xy) &= |(au + dbv)^2 - d(bu + av)^2| \\ &= |a^2u^2 + 2daubv + d^2b^2v^2 - db^2u^2 - 2daubv - da^2v^2| \\ &= |a^2u^2 + d^2b^2v^2 - db^2u^2 - da^2v^2| \\ &= |(a^2 - db^2)(u^2 - dv^2)| \\ &= N(x)N(y). \end{aligned}$$

Conjugates and norms are useful to compute ratios in the ring. Like $\mathbb{Z}$ is contained in $\mathbb{Q}$, so is $\mathbb{Z}[\epsilon]$ contained in $\mathbb{Q}[\epsilon]$, and $\mathbb{Q}[\epsilon]$ is a field. Indeed, $\mathbb{Q}[\epsilon]$ is a ring by the same reason $\mathbb{Z}[\epsilon]$ is. To prove it is a field, we need to show that every non-zero $x \in \mathbb{Q}[\epsilon]$ has an inverse $\frac{1}{x} \in \mathbb{Q}[\epsilon]$. Indeed, since $x \cdot \overline{x} \in \mathbb{Z}[\epsilon]$ one has $x \cdot \frac{\overline{x}}{x\overline{x}} = 1$ so $\frac{1}{x} = \frac{\overline{x}}{\pm N(x)}$ which belongs to $\mathbb{Q}[\epsilon]$. Correspondingly, $\mathbb{Z}[\epsilon]$ is called the ring of integers of $\mathbb{Q}[\epsilon]$ (for $\epsilon = \sqrt{d}$ or $\epsilon = \frac{1+\sqrt{d}}{2}$). The norm extends to $\mathbb{Q}[\epsilon]$ as well.

4 Divisibility in rings. Units and primes.

Now that we have somewhat established the properties of the quadratic rings, it is time to establish divisibility in them.

Definition 4.1. For x, y in a ring $R = \mathbb{Z}[\epsilon]$, we say that x divides y and write $x \mid y$, if there exists an element $z \in R$ such that $xz \in R$.

The following is a simple application of the multiplicativity of the norm.

Theorem 4.2. If $x, y \in R$ and $x \mid y$ then $N(x) \mid N(y)$.

Proof. If $x \mid y$ then $y = xz, z \in R$ so taking norms $N(y) = N(x)N(z)$ which means that $N(x) \mid N(y)$.

The converse is not necessarily true, as will be shown later.

Checking if $x \mid y$ in quadratic rings is easy. One needs to take the ratio $\frac{y}{x} = \pm\frac{y\overline{x}}{N(x)}$ which belongs to $\mathbb{Q}[\epsilon]$. So now we only need to check whether $\frac{y}{x}$ is in $\mathbb{Z}[\epsilon]$, which is done by simply writing $\frac{y}{x} \in \mathbb{Q}[\epsilon]$ as $a + b\sqrt{d}$ and checking if a and b are both integers. Keep in mind that a integer k divides a number $a + b\epsilon \in R$ if and only if k divides both of its parts a and b, because $\frac{a+b\epsilon}{k} = \frac{a}{k} + \frac{b}{k}\epsilon$. Now since $\frac{y}{x} = \pm\frac{y\overline{x}}{N(x)}$ we deduce the following property.

Theorem 4.3. If $x, y \in R$ then x divides y if and only if $N(x)$ divides $y\overline{x}$, that is $y\overline{x}$ has both components divisible by $N(x)$.

Note that if x has norm equal to 1, the above theorem tells us that x divides every other number y from R.

Definition 4.4. An element $u \in R$ is called a unit if it divides every element of R.

To check that u is a unit, it is enough to check that it is invertible, i.e. it has a inverse $\frac{1}{u}$ belonging to R. Indeed, if u is a unit then it must divide 1 so $\frac{1}{u}$ must belong to R. Conversely, if $\frac{1}{u} \in R$ then for any $x \in R$, $\frac{x}{u} = x \cdot \frac{1}{u} \in R$ so $u \mid x$. The following theorem characterizes the units of a quadratic ring.

Theorem 4.5. An element u of a quadratic ring is a unit if and only if its norm is 1.

Proof. If the norm of u is 1 then $u\overline{u} = \pm 1$ so $\frac{1}{u} = \pm\overline{u}$ so it belongs to the ring. Conversely, if u is a unit then $u \mid 1$ so $N(u) \mid N(1) = 1$ which implies that $N(u) = 1$.

The units are very important for understanding divisibility in rings. Indeed, since units divide every element of the ring, they are somehow irrelevant for purposes of divisibility and can always be ignored. We now understand why a proper divisibility theory for $\mathbb{R}$ and $\mathbb{Q}$, and for fields in general, cannot be constructed. As every element of a field has an inverse, it is a unit, which means every element divides each other so such a theory makes no sense. So the weakness of $\mathbb{Z}$ is in a sense its strength: the fact that in $\mathbb{Z}$ we can not always divide implies that the divisibility is not so simple as in $\mathbb{Q}$ or $\mathbb{R}$, thus a beautiful and rich theory can be constructed.

The units of $\mathbb{Z}$ are 1 and -1, but in $\mathbb{Z}[\epsilon]$ we can have more units.

Definition 4.6. Two numbers x and y are called associated (and we write $x \sim y$) if and only if they divide each other. Equivalently, their ratio must be a unit in R.

Indeed, $x \sim y$ if and only if $\frac{y}{x}$, $\frac{x}{y}$ are both in R i.e. $\frac{x}{y}$ is in R and has an inverse, so is a unit. Clearly, units are precisely numbers associated with 1. Often for purposes of divisibility associated numbers are indistinguishable.

Definition 4.7. A number $\pi \in \mathbb{R}$ is called a prime if it is only divisible by units and numbers associated with it. Equivalently, π can not be written as $\pi_1\pi_2$ where π_1, π_2 are not units.

This definition mirrors the definition in $\mathbb{Z}$. A little remark must be made: according to the definition, -2 should also be a prime in $\mathbb{Z}$, which contradicts our knowledge that 2 is a prime and -2 is not. However, this is only because we are used to constructing divisibility in $\mathbb{N}$, the positive part of $\mathbb{Z}$. In the general case, we cannot perform this distinction (what are the positive numbers in $\mathbb{Z}[\sqrt{-1}]$?), so we must deal with the units, too. In fact, every number associated with a prime is also a prime, which means we somehow have multiple copies of every prime. But this is not a huge problem since we can ignore units when we need to.

Now we have given the definition of a prime, but do we even know they exist, and more importantly, do we know that we can factorize each number into primes, like in $\mathbb{Z}$? The answer is positive in both cases, and relies on norm.

Assume we have a number $r \in R$. If r is a prime, we are done. If not, then $r = r_1r_2$ and none of the r_1, r_2 is a unit so $N(r_1), N(r_2) > 1$. The equality $N(r_1)N(r_2) = N(r)$ tells that $N(r_1), N(r_2) < N(r)$. Now we work with r_1, r_2, and if they are not primes again we decompose them into factors. As we see that the norm of the factors decreases, we cannot continue the process indefinitely, so we must eventually stop, which gives us a decomposition of r into factors that cannot be decomposed further, i.e. into primes.

We have almost constructed a divisibility theory for rings by simply copying the theory for $\mathbb{Z}$. There is one important question left: is the factorization into primes unique? Such rings are called Unique Factorization Domains (UFD's) and we are considering them for the purposes of these notes.

We must note that decomposition into primes must be unique up to units, since for example $pq = (-p)(-q)$ are two different decompositions of the same number. But if we decide to consider associated primes as signifying the same prime, we get rid of this problem. So the question is: given two decompositions of the same number into primes, are the primes entering the two decompositions pairwise associated?

Unfortunately, the general answer is no, and this is a huge deception. In $\mathbb{Z}[\sqrt{-5}]$, $6 = 2 \cdot 3 = (1+\sqrt{-5})(1-\sqrt{-5})$. It turns out that these four numbers are all primes, and none them is associated with another.

Indeed, for example 2 has norm 4 so if it would not be a prime it would be $\pi_1\pi_2$ with $N(\pi_1), N(\pi_2) > 1$ which can only happen when $N(\pi_1) = N(\pi_2) = 2$. But there can be no numbers of norm 2 in $\mathbb{Z}[\sqrt{-5}]$: if $N(x + \sqrt{-5}y) = 2$ then $x^2 + 5y^2 = 2$ which is clearly impossible since than we must have $y = 0$ (otherwise

$5y^2 > 2$) and so $x^2 = 2$, which has no solutions in $\mathbb{Z}$. We deduce similarly that $3, 1+\sqrt{-5}, 1-\sqrt{-5}$ are primes, and they are not associated with each other since they have different norms (well, $1+\sqrt{-5}$ and $1-\sqrt{-5}$ have the same norm, but we only need to compare factors from the left-hand side to factors from the right-hand side).

This failure is very disappointing, and now we know we cannot construct a proper divisibility theory for all quadratic rings. However for some small values of d we still can achieve success, but we need another tool borrowed from $\mathbb{Z}$.

5 The Euclidean Algorithm

In $\mathbb{Z}$ (well, actually in $\mathbb{N}$, but we can extend the method for $\mathbb{Z}$), we have the Euclidean algorithm, that gives us a way to construct the greatest common divisor of any two numbers as a linear combination of them. If you have a good memory (I don't but I still remember), you may recall that the ability to find the common divisor of any two numbers was enough to prove that prime factorization is unique. Indeed, if we would be able to construct an Euclidean Algorithm in $\mathbb{Z}[\sqrt{-5}]$, then we would have $1 = 2a + (1+\sqrt{5})b$ (as their greatest common divisor is clearly 1), but this is impossible as the norm of $2a + (1+\sqrt{5})b$ is $(2a+b)^2 + 5b^2$ which is always even. So now we set to formulate an Euclidean Algorithm for at least some quadratic rings R, unless we want to fail completely and close the topic.

The Euclidean Algorithm was based on the remainder: for any $a, b \in \mathbb{Z}$, $b \neq 0$, there exists a number $q \in \mathbb{Z}$ and r with $0 < |r| < |b|$ such that $a = bq + r$. Then, it replaced the pair (a, b) with the pair (b, r) and continued doing the same thing until it reached 0 (and it did, because the absolute values decreased), and the other number of the final pair was exactly the greatest common divisor.

In rings R, the absolute value is replaced by norm. So in order to concoct an Euclidean Algorithm, we need to be able to prove the following fact:

Theorem 5.1. If $a, b \in R$ and R is a quadratic ring, then for $b \neq 0$ there exist $q, r \in R$ such that $N(r) < N(b)$ such that $a = bq + r$.

If this theorem is true, then the Euclidean Algorithm will work. Unfortunately, the example of $\mathbb{Z}[\sqrt{-5}]$ shows that theorem 5.1. is not true for all rings, but as we will see, it is true in some simple cases.

The condition $a = bq + r$ is equivalent to $\frac{a}{b} = q + \frac{r}{b}$ and $N(\frac{r}{b}) < 1$. If $R = \mathbb{Z}[\epsilon]$ then $\frac{a}{b} \in \mathbb{Q}[\epsilon]$. The theorem 5.1. now means that for every $x \in \mathbb{Q}[\epsilon]$ there exists a number $q \in \mathbb{R}$ such that $N(x - q) < 1$.

Let $x = u + v\epsilon$. It is reasonable to set $q = a + b\epsilon$ where a is the closest integer to u, b is the closest integer to v. We know that $|a - u| \leq \frac{1}{2}, |b - v| \leq \frac{1}{2}$ so $x - q$

has both components of absolute value at most 1. It is now needed to establish for which ϵ this implies that $N((a-u)+(b-v)\epsilon) < 1$.

Recall that we considered $\epsilon = \sqrt{d}$ for $d \equiv 2, 3 \pmod 4$ and $\epsilon = \frac{1+\sqrt{d}}{2}$ for $d \equiv 1 \pmod 4$.

In the first case,

$$N((a-u)+(b-v)\epsilon) = |(a-u)^2 - d(b-v)^2|$$

and $(a-u)^2 \leq \frac{1}{4}, (b-v)^2 \leq \frac{1}{4}$.

For d negative, we conclude $|(a-u)^2 - d(b-v)^2| \leq \frac{1}{4} + \frac{|d|}{4} = \frac{|d|+1}{4}$. Thus $\frac{|d|+1}{4} < 1$ which means $|d| < 3$ so $d \in \{-2, -1\}$.

For d positive, we have $|(a-u)^2 - d(b-v)^2| \leq \frac{d}{4}$ so we need $\frac{d}{4} < 1$ so $d \in \{2, 3\}$ ($d = 1$ is a perfect square).

In the second case,

$$N((a-u)+(b-v)\epsilon) = |(a-u)^2 + (a-u)(b-v) + \frac{1-d}{4}(b-v)^2|.$$

For d negative we have $|(a-u)^2 + (a-u)(b-v) + \frac{1-d}{4}(b-v)^2| \leq \frac{1}{4} + \frac{1}{4} + \frac{1-d}{4} \cdot \frac{1}{4} = \frac{9-d}{16}$ and we need $\frac{9-d}{16} < 1$ i.e. $d > -7$ or $d = -3$.

For d positive we have $|(a-u)^2 + (a-u)(b-v) + \frac{1-d}{4}(b-v)^2| \leq \frac{d-1}{16} + \frac{1}{4} + \frac{1}{4} = \frac{d+7}{16}$ and we need $d + 7 < 16$ which gives $d = 5$.

For the values of d we have just established, the Euclidean Algorithm works. Such rings are called Euclidean Domains.

Theorem 5.2. $\mathbb{Z}[\sqrt{-1}], \mathbb{Z}[\sqrt{-2}], \mathbb{Z}[\sqrt{2}], \mathbb{Z}[\sqrt{3}], \mathbb{Z}[\frac{1+\sqrt{-3}}{2}], \mathbb{Z}[\frac{1+\sqrt{5}}{2}]$ are Euclidean Domains.

In Euclidean domains, just like in $\mathbb{Z}$, the Euclid lemma is true:

Theorem 5.3. (Euclid's Lemma) If R is an Euclidean domain and $a, b \in R$, then the greatest common divisor of a, b exists and can be written as $sa + tb$ for $s, t \in R$. The greatest common divisor of a and b is denoted by (a, b) or $gcd(a, b)$.

The proof of this theorem is just like in the integer case, and we will not repeat it here.

The greatest common divisor of a and b can be defined as the number c such that $c \mid a$, $c \mid b$ and if $c' \mid a$, $c' \mid b$ then $c' \mid c$. It is clear that the greatest common divisor is determined up to a unit: if c is the greatest common divisor, so is any number associated with it.

If two numbers have their greatest common divisor equal to 1 (i.e. the only common factors are units) then we say these two numbers are relatively prime (coprime).

The following fact is again copied from $\mathbb{Z}$:

Theorem 5.4. If $a \mid bc$ and a is relatively prime to b, then $a \mid c$.

Proof. If $(a, b) = 1$ then $ax + by = 1$ for some $x, y \in R$. Multiplying by c we get $c = acx + bcy$. Since acx and bcy are divisible by a, we deduce c also is.

If π is a prime then its factors are only units and numbers associated to it. It follows that every number is either divisible by π, or is relatively prime to it. Therefore theorem 5.4. applied to primes says:

Theorem 5.5. If π is a prime and $\pi \mid ab$, then $\pi \mid a$ or $\pi \mid b$.

Again those with good memory will remember that theorem 5.5. was the main ingredient of the proof of the unique factorization theorem for $\mathbb{Z}$. Transferring the argument for the Euclidean Domains, one gets the following:

Theorem 5.6. If R is an Euclidean Domain, then the prime factorization of every number is unique, up to units. That is, if $\pi_1\pi_2 \dots \pi_m = \pi'_1\pi'_2 \dots \pi'_n$ and $\pi_1, \dots, \pi_m, \pi'_1, \dots \pi'_n$ are primes, then $m = n$ and we can reorder the primes such that $\pi_i \sim \pi'_i$ $(i = 1, 2, \dots, m)$.

Particularly, the quadratic rings mentioned in theorem 5.3. are UFD's.

6 Classification of primes

Having established the unique decomposition into primes in some quadratic rings R, we should ask ourselves what are the primes in R.

Firstly, we need some convention. From now on, "primes" will signify prime numbers in R. The primes in $\mathbb{Z}$ will be called "integer primes".

It is clear that a number is a prime if and only if its conjugate is prime. Indeed, any factorization of x can be transformed into a factorization of $\overline{x}$ by conjugation, and vice-versa.

The norm gives us an invaluable insight.

Theorem 6.1. If $N(\pi)$ is prime (in $\mathbb{Z}$) then π is a prime (in R).

Proof. If $\pi = xy$ for non-units x, y then $N(\pi) = N(x)N(y)$ and $N(x), N(y) > 1$, which contradicts the assumption that $N(\pi)$ is prime.

We see that if the norm of a number is prime integer, then the number itself is a prime. The converse does not hold, as the example of 3 in $\mathbb{Z}[\sqrt{-1}]$ shows. Indeed, $N(3) = 9 = 3^2$ so if $3 = \pi_1\pi_2$ for π_1, π_2 not units, then we must have $N(\pi_1) = N(\pi_2) = 3$, which is impossible since the equation $a^2 + b^2 = 3$ has no solution in integers. However, the norm of a prime number must not have too many divisors.

Firstly, if π is a prime then it must divide an integer prime. Indeed, $\pi \mid N(\pi)$ which is an integer, so is a product of integer primes. Theorem 5.5 now implies that π divides one of these primes, say p. Then $N(\pi) \mid N(p) = p^2$, so $N(\pi) = p$ or $N(\pi) = p^2$.

In the first case, $\pi\overline{\pi} = p$, so p is the product of two conjugate primes of norm p. In the second case $N(\pi) = N(p)$ and as $\pi \mid p$ it means $\pi \sim p$ so p is a prime in R. We must now distinguish these two cases. Let p be odd.

Assume the first case holds. Then $p = N(\pi) = a^2 - db^2$ where a, b are either integers or half-integers so $4p = (2a)^2 - d(2b)^2$ which means that d is a perfect square modulo p.

Assume the second case holds. We claim that d is not a square modulo p. Indeed, if it were then p would divide $x^2 - d = (x - \sqrt{d})(x + \sqrt{d})$ for some integer x, contradicting theorem 5.5 because p divides neither $x + \sqrt{d}$ nor $x - \sqrt{d}$.

We thus have established that for p odd, the first case holds when d is a square residue modulo p, and the second case holds when d is not a square residue modulo p.

The results obtained can be encoded in the following theorem:

Theorem 6.2. In a quadratic ring R, all primes must have divide an integer prime p and so their norm is either p or p^2. If p is odd and d a perfect square modulo p, then p is the product of two conjugate primes of norm p. If p is odd and d is not a perfect square modulo p, then p is a prime in R.

We are left to consider the case $p = 2$. Let's analyze the rings one by one:

• In $\mathbb{Z}[\frac{1+\sqrt{-3}}{2}]$, 2 is a prime. Indeed, if it was not then we would have $2 = (a + b\epsilon)(a - b\epsilon) = a^2 - ab + b^2$ for $a, b \in \mathbb{Z}$. However $a^2 - ab + b^2$ has no solutions, as easily seen modulo 4.

• In $\mathbb{Z}[\sqrt{-2}]$, 2 is associated to the square of a prime, as $2 = -\sqrt{-2}^2$.

• In $\mathbb{Z}[\sqrt{-1}]$, 2 is associated to the square of a prime, as

$$2 = (1 + \sqrt{-1})(1 - \sqrt{-1}) = -\sqrt{-1}(1 + \sqrt{-1})^2.$$

• In $\mathbb{Z}[\sqrt{2}]$, 2 is the square of the prime $\sqrt{2}$.

• In $\mathbb{Z}[\sqrt{3}]$, 2 is the product of two non-associated primes $\sqrt{3}+1$ and $\sqrt{3}-1$.

• In $\mathbb{Z}[\frac{1+\sqrt{5}}{2}]$, 2 is prime. If it wasn't, then we would have $2=(a+b\epsilon)(a-b\epsilon)=a^2-ab-b^2$, and this equation has no solution modulo 4.

It should be noted that unless $d=p$ or $p=2$ and $R=\mathbb{Z}[\sqrt{3}],\mathbb{Z}[\sqrt{-1}]$, then the two conjugated primes dividing p are not associated (if p is the product of two primes). Indeed, if $p=\pi\overline{\pi}$ and $\pi\sim\overline{\pi}$ then $\pi^2\sim p$ and hence

$$p \mid 4(a+b\sqrt{d})^2=(2a)^2+d(2b)^2+8abd,$$

which implies $p \mid 4ab, p \mid 4a^2+d4b^2$ so $p \mid 2a, p \mid 2b$. This is impossible because $(2a)^2-d(2b)^2=4p$ is not divisible by p^2.

Definition 6.3. If an integer prime decomposes as the product of two primes in R, we say p splits in R. If the two primes are associated, we say p ramifies in R.

7 The Chinese Remainder Theorem. The Residue Class Field. Fermat and Euler.

The congruence can be naturally defined for Unique Factorization Domains: one says that $x\equiv y \pmod{z}$ if and only if $z \mid x-y$. The congruence clearly inherits all the properties of its elementary counterpart.

Theorem 7.1. (The Chinese Remainder Theorem) Given pairwise coprime numbers $b_1,b_2,\ldots,b_n$ and any numbers $a_1,a_2,\ldots,a_n$ in R there exists a number $x\in R$ such that $x\equiv a_i \pmod{b_i}$. Any two such numbers are congruent modulo $b_1b_2\cdots b_n$.

We will not prove this theorem, since it is a shameless imitation of the proof for $\mathbb{Z}$.

For a number $p\in\mathbb{Z}$, we can define $\mathbb{Z}/p\mathbb{Z}$ be the set of all residue classes modulo p. It is naturally a ring, and even a field if p is prime, and has $|p|$ elements. Similarly, for a number π in R, we can define $R/\pi R$ be the set of all residues modulo π. We will write $\hat{x}$ for the class of residues modulo π represented by x.

Theorem 7.2. For a number π, the residue class $R/\pi R$ is a finite ring containing $N(\pi)$ elements. Moreover, $R/\pi R$ is a field if and only if π is prime.

We will prove this problem in several steps. The easiest to prove is that $R/\pi R$ is finite. Indeed, as $\pi \mid N(\pi)$, we conclude that two numbers $a_1+b_1\epsilon$ and $a_2+b_2\epsilon$ that satisfy $a_1\equiv a_2 \pmod{N(\pi)}$ and $b_1\equiv b_2 \pmod{N(\pi)}$ will also be congruent modulo π. Since there are $N(\pi)^2$ pairs of integer residues modulo $N(\pi)$, it follows

that $R/\pi R$ has at most $N(\pi)^2$ elements, so is finite. Note that instead of $N(\pi)$ one can take any integer N divisible by π, and one gets the bound N^2.

Next, if π is not prime write $\pi = \pi_1\pi_2$ for non-units π_1, π_2. We have the equality $\hat{\pi_1}\hat{\pi_2} = \hat{0}$. Since $\hat{\pi_1}, \hat{\pi_2}$ are non-zero, this equality implies that $R/\pi R$ is not a field, since then $\hat{\pi_1}, \hat{\pi_2}$ are not invertible.

If π is a prime then $R/\pi R$ is a field, and the proof is the same as in the elementary case. If $\hat{a}$ is a non-zero residue modulo π, then $(a, \pi) = 1$ hence by Euclid's lemma there exist b, c such that $ab + c\pi = 1$. Thus $\hat{a}\hat{b} = \hat{1}$ so $\hat{a}$ is invertible.

The hardest part is to prove that $R/\pi R$ has exactly $N(\pi)$ elements. We will do this by induction on the number of prime factors of π.

First, assume that π is a prime. As we know $\pi \mid p$ for some integer prime p and we either have $N(\pi) = p$ or $N(\pi) = p^2$ and $\pi \sim p$. In the first case, $\pi = a + b\epsilon$ hence $b\epsilon \equiv a \pmod{\pi}$. Since b is invertible modulo p, we conclude that ϵ is congruent to ab_1 modulo π, where b_1 is chosen to satisfy the congruence $bb_1 \equiv 1 \pmod{p}$. We conclude that ϵ is congruent to an integer modulo π and from here every element of R is congruent to an integer modulo π. Since there are only p classes in $\mathbb{Z}$ modulo p and $\pi \mid p$, we conclude that the residue class field has at most p elements. Now, the classes $\hat{0}, \hat{1}, \ldots, \hat{p-1}$ are all distinct, since $\hat{i} = \hat{j}$ implies $\pi \mid i - j$ so by passing to norm we get $p \mid (i-j)^2$, impossible. Thus there are exactly p classes, which proves the claim in this case. In the second case, we have $\pi \equiv p$ and clearly that $R/\pi R$ is the same as R/pR. Now as p is an integer, we conclude that R/pR has at most p^2 elements. It is enough now to exhibit p^2 distinct classes to conclude the claim in this case too. Indeed, consider the p^2 numbers $i + j\epsilon$ where $i, j \in \{0, 1, \ldots, p-1\}$. They all give distinct residues, for if $i_1 + j_1\epsilon \equiv i_2 + j_2\epsilon \pmod{p}$ then $p \mid (i_1 - i_2) + \epsilon(j_1 - j_2)$ so $p \mid i_1 - i_2, p \mid j_1 - j_2$, which is impossible. Thus R/pR has p^2 elements, as desired.

The case π is prime is the base. Now let's perform the induction step. Assume that $\pi = \pi_1 x$ where π_1 is a prime, so the claim is proven for π_1 and for x. Consider $b_1, \ldots, b_{N(\pi_1)}$ be a set of representatives modulo π_1 and $c_1, \ldots, c_{N(x)}$ a set of representatives modulo x. We claim that the numbers $b_i + \pi c_j$ are a complete set of representatives modulo $\pi_1 x$. Since the number of such elements is $N(\pi_1)N(x) = N(\pi)$, this will finish the proof the theorem. Indeed, for every $r \in R$, $r \equiv b_i \pmod{\pi_1}$ for some i so $r - b_i = \pi_1 s$ for $s \in R$. Then $s \equiv c_j \pmod{x}$ for some j so $s - c_j = xt$ for some $t \in R$. Then $r - (b_i + \pi_1 c_j) = \pi_1 s - \pi_1 c_j = \pi_1 xt = \pi t$ hence $r \equiv b_1 + \pi_1 c_j \pmod{\pi}$. We are left to prove that $b_i + \pi_1 c_j$ are pairwise non-congruent modulo π. Indeed, if $b_i + \pi_1 c_j \equiv b_k + \pi_1 c_l \pmod{\pi}$ then $b_i \equiv b_k \pmod{\pi_1}$ so $i = k$. Then $(b_i + \pi_1 c_j) - (b_k + \pi_2 c_l) = \pi_1(c_j - c_l)$. Since it must be divisible by $\pi = \pi_1 x$, $c_j - c_l$ must be divisible by x, possible only when $j = l$ i.e. (i, j) and (k, l) are the same pair. The proof is finished.

Theorem 7.3. (Fermat) Given a prime π the following congruence holds for all a:

$$a^{N(\pi)} \equiv a \pmod{\pi}.$$

Proof. Again, it is a reproduction of the elementary version. If a is divisible by π, the congruence is clear. If a is not, then it is coprime to π. If $\hat{x_1}, \hat{x_2}, \ldots, x_{\hat{N(\pi)-1}}$ are the non-zero residues modulo π, then $\hat{a}\hat{x_1}, \hat{a}\hat{x_2}, \ldots, \hat{a}x_{\hat{N(\pi)-1}}$ must also be distinct non-zero residues, so they must be a permutation of $\hat{x_1}, \hat{x_2}, \ldots, x_{\hat{N(\pi)-1}}$. Taking the product one gets $\hat{x_1}\hat{x_2}\cdots x_{\hat{N(\pi)-1}} = \hat{a}\hat{x_1}\hat{a}\hat{x_2}\cdots\hat{a}x_{\hat{N(\pi)-1}}$. Now simplifying the invertible elements $\hat{x_1}, \hat{x_2}, \ldots, x_{\hat{N(\pi)-1}}$ one get $\hat{1} = \hat{a}^{N(\pi)-1}$. Thus $a^{N(\pi)-1} \equiv 1 \pmod{\pi}$ or $a^{N(\pi)} \equiv a \pmod{\pi}$.

For each number π, we can denote by $\phi(\pi)$ the number of residues in $R/\pi R$ that are coprime to π (which are invertible). Fermat's Theorem generalizes to

Theorem 7.4. (Euler) The following congruence holds for a coprime to π:

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

Like in the previous theorem, the proof is a mere adaptation of the elementary version, and we will not repeat it here.

Now, $\phi(\pi)$ can be explicitly computed, and the proof is again as unoriginal as possible.

Theorem 7.5. If $\pi \equiv \prod_{i=1}^{m} {\pi_i}^{a_i}$ where $\pi_1, \pi_2, \ldots, \pi_m$ are distinct primes, then

$$\phi(\pi) = \prod_{i=1}^{m} \left(N(\pi_i)^{a_i} - N(\pi_i)^{a_i-1}\right) = N(\pi)\prod_{i=1}^{m}\frac{N(\pi_i)-1}{N(\pi_i)}.$$

Proof. The residue of any number r modulo π is uniquely determined by the residues r_i of this number modulo $\pi_i^{a_i}$, according to the Chinese Remainder Theorem. In order for r to be coprime to π, it is necessary and sufficient that every of the r_i be not divisible by π_i. Now there are $N(\pi_i)^{a_i}$ residues modulo $\pi_i^{a_i}$. Of these, a fraction of $\frac{1}{N(\pi_i)}$ is divisible by π_i (prove it yourself!), so there are $(1 - \frac{1}{N(\pi_i)})N(\pi_i)^k = N(\pi_i)^k - N(\pi_i)^{k-1}$ admissible residues for r_i. Thus there are $\prod_{i=1}^{m}\left(N(\pi_i)^k - N(\pi_i^{k-1})\right)$ admissible m-tuples $(r_1, r_2, \ldots, r_m)$. Each of these m-tuples gives rise to a unique residue modulo π, by the Chinese Remainder Theorem, and this finishes the proof.

Theorem 7.6. (Wilson) If π is a prime, then the product of all non-zero residues modulo π is congruent to -1 modulo π.

The proof is again, an IQ-free copy-paste.

8 Numbers with bounded norm. Units. Pell's Equation.

We have mentioned before that numbers with the same norm are not necessarily associated. However, we can prove that there are finitely many non-associated numbers with a given norm.

Theorem 8.1 For any positive integer M, there are only finitely many pairwise non-associated numbers with norm M.

Proof 1. If a number has norm M then it divides M. Writing M as a product of primes, it is clear that one can select finitely many divisors (up to units). Namely, if $M \sim \prod_{i=1}^{m} {\pi_i}^{a_i}$, every divisor of it is associated to a number of the form $\prod_{i=1}^{m} {p_i}^{b_i}$ for $b_i \in \{0, 1, \ldots, a_i\}$, and there are only $\prod_{i=1}^{m}(a_i+1)$ such choices of $(b_1, b_2, \ldots, b_m)$.

Proof 2. We claim among any M^2+1 numbers with norm M two are associated. Indeed, assume $x_i + y_i\epsilon, i = 1, 2, \ldots, M^2+1$ are numbers with norm M. Since there are only M^2 pairs of residues modulo M^2, by the pigeonhole principle there exist two numbers $u = x_k + y_k\epsilon, v = x_l + y_l\epsilon$ such that $M \mid x_k - x_l$, $M \mid y_k - y_l$. Then $M \mid u - v$ and since $u, v \mid M$ we conclude $u \mid v$ and $v \mid u$ so $u \sim v$, as desired.

Note that the bound obtained in the first proof is much smaller then the bound obtained in the second.

As an immediate corollary, one can prove

Theorem 8.2. For any $M > 0$, there are only finitely many pairwise non-associated numbers with norm less than M.

Now, since the ratio of two numbers is always a unit, it would be helpful to know more about the units of the ring.

For $R = \mathbb{Z}[\sqrt{-1}], \mathbb{Z}[\sqrt{-2}]$ or $R = \mathbb{Z}[\frac{1+\sqrt{-3}}{2}]$, the units can be easily listed.

• For $\mathbb{Z}[\sqrt{-1}]$, $N(a+b\sqrt{-1}) = 1$ means $a^2+b^2 = 1$ possible only for $a = \pm 1, b = 0$ or $a = 0, b = \pm 1$ so $\pm 1, \pm\sqrt{-1}$ are the only units.

• For $\mathbb{Z}[\sqrt{-2}]$, one similarly gets the equation $a^2 + 2b^2 = 1$ possible only for $a = \pm 1$ so the only units in this ring are 1 and -1.

• For $\mathbb{Z}[\frac{1+\sqrt{-3}}{2}]$, one gets the equation $x^2+y^2+xy = 1$, i.e. $3(x+y)^2+(x-y)^2 = 4$, which only has $x = \pm 1, y = 0$, $x = 0, y = \pm 1$ and $x = 1, y = -1$ or $x = 1, y = -1$ as solutions. So the units are ± 1 and $\frac{\pm 1 \pm \sqrt{-3}}{2}$. These are the sixth roots of unity.

For other rings this method does not work. Indeed, for $d > 0$, the norm of $x + y\sqrt{d}$ is $|x^2 - dy^2|$ and now we cannot conclude that x, y are bounded. In fact,

we have infinitely many solutions to the equation $|x^2 - dy^2| = 1$ and we will use theorem 8.2. to prove it.

This time we do not assume that $d > 0$ is square-free and neither that $R = \mathbb{Z}[\sqrt{d}]$ is a UFD. Indeed, we will only use theorem 8.2 (and its second proof does not require UFD's) and only initial considerations from section 4.

Consider the numbers $\{\sqrt{d}\}, \{2\sqrt{d}\}, \ldots, \{(N+1)\sqrt{d}\} \in (0,1)$. By the pigeonhole principle there are two of them which are at most $\frac{1}{N}$ apart, say $|\{i\sqrt{d}\} - \{j\sqrt{d}\}| \leq \frac{1}{N}$. Since $\{i\sqrt{d}\} = i\sqrt{d} - k$ for some integer k and $\{j\sqrt{d}\} = j\sqrt{d} - l$ for some integer l, we deduce $|(i-j)\sqrt{d} - (k-l)| \leq \frac{1}{N}$ i.e. we have x_1, y_1 such that $|x_1 - y_1\sqrt{d}| < \frac{1}{N}$. We also deduce $|x_1| \leq \sqrt{d}N$, $|y_1| \leq N$ so

$$|x_1^2 - dy_1^2| = |x_1 - y_1\sqrt{d}||x_1 + \sqrt{d}N| \leq \frac{1}{N} \cdot (\sqrt{d}N + \sqrt{d}N) < 2\sqrt{d}.$$

We thus have found a number $z_1 = x_1 - y_1\sqrt{d}$ in the ring R with norm at most $2\sqrt{d}$. Now we can pick up a larger N' and apply the method to find another x_2, y_2 such that $z_2 = x_2 - y_2\sqrt{d}$ has again norm at most $2\sqrt{d}$. Moreover, if we choose N' big enough we will have $x_1 - y_1\sqrt{d} \neq x_2 - y_2\sqrt{d}$. Indeed, since d is irrational, $|x_1 - y_1\sqrt{d}| > 0$, so taking $N' > \frac{1}{|x_1 - y_1\sqrt{d}|}$, we must have $|x_2 - y_2\sqrt{d}| < |x_1 - y_1\sqrt{d}|$ which implies the result. Continuing, we find an infinite number of numbers $z_1, z_2, \ldots$ which have all norm at most $2\sqrt{d}$. But theorem 8.2. guarantees there are only finitely many of them which are pairwise non-associated. Let $z_1, z_2, \ldots, z_k$ be a maximal set of pairwise non-associated members of the sequence. Then every z_i is associated to one of $z_1, z_2, \ldots, z_k$ which means that infinitely many numbers z_i are associated with, say, z_1. This means infinitely many of the $\frac{z_i}{z_1}$ are units, and since they are distinct we have obtained infinitely many units. This means that there are infinitely many pairs of numbers a, b such that $|a^2 - db^2| = 1$. Then $a^2 - db^2 = 1$ or $a^2 - db^2 = -1$ but in any case $(a^2 - db^2)^2 = (a^2 + db^2) - d(2ab)^2 = 1$. Therefore we have proved the following theorem:

Theorem 8.3. If $d > 0$ is not a perfect square, then Pell's Equation $x^2 - dy^2 = 1$ has infinitely many solutions in integers.

9 Applications and Examples

It is time to test in battle all the machinery I've spent so much effort to write and you've spent so much effort to read. You are encouraged to seek solutions without UFD's, since I hope this will enhance your respect for them.

Example 9.1. Solve the equation $x^2 + 2y^2 = 3z^2$ in integers.

Solution. If is clear that if (x, y, z) have a common divisor d, then $(\frac{x}{d}, \frac{y}{d}, \frac{z}{d})$ is also a solution, hence it suffices to consider the case $gcd(x, y, z) = 1$. Then we

can conclude $(x,y)=1$. Otherwise, if $p \mid x$, $p \mid y$ for p prime, then $p^2 \mid x^2+2y^2$ so $p^2 \mid 3z^2$ hence $p \mid z$ and $(x,y,z) \neq 1$. Similarly, one can prove $(x,z)=1$ and $(y,z)=1$.

Now we can factor the left-hand side in $R=\mathbb{Z}[\sqrt{-2}]$ and we get $(x+\sqrt{-2}y)(x-\sqrt{-2}y)=3z^2$. Set $u=x+y\sqrt{-2}, v=x-y\sqrt{-2}$. We prove u,v are coprime. Indeed, let π be a prime divisor of u and v. Then $\pi \mid v+u=2x$ and $\pi \mid u-v=2\sqrt{-2}y$. As x,y are coprime in $\mathbb{Z}$ then they are coprime in R, too (because their norms are coprime). It follows that π must divide either $2\sqrt{-2} \sim (\sqrt{-2})^3$. We can only have $\pi \sim \sqrt{-2}$ which implies that $\pi \mid uv=3z^2$ so $\pi \mid z$. Hence $N(\pi) \mid z^2$ thus z is even and so must be x, violating the condition $(x,z)=1$. Thus u,v must be coprime.

Now their product is $3z^2=(1+\sqrt{-2})(1-\sqrt{-2})z^2$. Since $1+\sqrt{-2}, 1-\sqrt{-2}$ are primes, we must conclude that $u=\pm 3r^2, v=\pm s^2$ or $u=\pm(1+\sqrt{-2})r^2, v=\pm(1-\sqrt{-2})s^2$ (since the only units in $\mathbb{Z}[\sqrt{-2}]$ are $1,-1$) or two other analogous cases. The first case is impossible since then $3 \mid u$ so $3 \mid \overline{u}=v$ hence u,v are not coprime.

Assume $u=(1+\sqrt{-2})r^2, v=(1-\sqrt{-2})s^2$. If $r=a+b\sqrt{-2}$ we get

$$\begin{aligned} x+y\sqrt{-2} &= (1+\sqrt{-2})(a+b\sqrt{-2})^2 \\ &= (1+\sqrt{-2})(a^2-2b^2+2ab\sqrt{-2}) \\ &= a^2-2b^2-4ab+(a^2-2b^2+2ab)\sqrt{2}. \end{aligned}$$

We then deduce $x=a^2-2b^2-4ab, y=a^2-2b^2+2ab$ hence we compute $z=\pm(a^2+2b^2)$. The other cases are similar and lead, essentially, to the same solution (up to the sign). Now recalling that we divided in the beginning all x,y,z by their common divisor d, we get $x=\pm d(a^2-2b^2-4ab), y=\pm d(a^2-2b^2+2ab), z=\pm d(a^2+2b^2)$. Also note that a,b must be coprime and a must be odd, in order for $a^2-2b^2-4ab, a^2+2b^2, a^2-2b^2+2ab$ to be coprime.

Example 9.2. (IMO Shortlist, 1988) Let x,y,z be positive integers such that $xy=z^2+1$. Prove that there exist integers a,b,c,d such that $x=a^2+b^2$, $y=c^2+d^2$.

Solution. In $\mathbb{Z}[\sqrt{-1}]$, $xy=(z+i)(z+i)$ (we set $\sqrt{-1}=i$, a thing we should have done long ago). Every prime π dividing z^2+1 must divide $z+i$ or $z-i$, so π cannot be an integer. Otherwise, π would have to divide the imaginary part of $z \pm i$ which is ± 1. Now if $\pi \mid x$, say, then by conjugating we deduce $\overline{\pi} \mid x$ and $\pi \neq \overline{\pi}$ (because no prime ramifies in $\mathbb{Z}[i]$). Therefore all prime factors of x split into pairs of conjugates, and similarly for y. We then deduce that $x \sim \prod_{k=1}^m \pi_k \overline{\pi_k}$ where π_k are non-integer primes. If we denote $a+bi=\prod_{k=1}^m \pi_k$ we see that $x \sim (a+bi)\overline{(a+bi)}=a^2+b^2$. So $\frac{x}{a^2+b^2}$ is a unit. Since in our case the units are $1,-1,i,-i$ but we know that x is a positive integer, we deduce that $x=a^2+b^2$. Similarly $y=c^2+d^2$.

Example 9.3. (Moldova TST, 2004) Let n be a positive integer and let $0 < a < c \leq d < b$ be positive integers such that $n = a^2 + b^2 = c^2 + d^2$. Show that n is not a prime.

Solution. Assume n would be a prime. Then $N(a + bi)$ is a prime integer so $a+bi$ is prime in $\mathbb{Z}[i]$, and similarly $a-bi$ is prime. Thus $(a+bi)(a-bi)$ is the prime factorization of n in $\mathbb{Z}[i]$. Analogously, $(c + di)(c - di)$ is the prime factorization of n in $\mathbb{Z}[i]$. Since the factorization is unique, we conclude that $a + bi \equiv c + di$ or $a + bi \equiv c - di$. Therefore $a + bi = \pm(c \pm di)$ or $a + bi = \pm i(c \pm di)$. We will then get $a = \pm c, b = \pm d$ or $a = \pm d, b = \pm c$. Since the numbers are positive, we have $a = c, b = d$ or $a = d, b = c$, which contradicts the statement of the problem.

Example 9.4. Let p be a prime number such that $\left(\frac{-2}{p}\right) = 1$. Show that there exist integers x, y such that $p = x^2 + 2y^2$.

Solution. By theorem 6.2 $\mathbb{Z}[\sqrt{-2}]$ contains a prime $x + \sqrt{-2}y$ of norm p. Then $N(x + \sqrt{-2}y) = p$ means $x^2 + 2y^2 = p$, as desired.

Example 9.5. Let $(F_n)_{n\in\mathbb{N}}$ be the Fibonacci sequence: $F_1 = 1, F_2 = 1, F_{n+2} = F_{n+1} + F_n$. Prove that for every m there exists a number in the sequence divisible by m.

Solution. Let $u = \frac{1+\sqrt{5}}{2}$. By Binet's formula,

$$F_n = \frac{(\frac{1+\sqrt{5}}{2})^n - (\frac{1-\sqrt{5}}{2})^n}{\sqrt{5}} = \frac{u^n - (-\frac{1}{u})^n}{\sqrt{5}}.$$

To avoid sign matters, assume that $n = 2k$ is even. Then $F_n = \frac{u^n - \frac{1}{u^n}}{\sqrt{5}} = \frac{u^{2n}-1}{\sqrt{5}u^n}$.

In the UFD $\mathbb{Z}[u]$, u is a unit. In order to have $m \mid F_n$ in $\mathbb{Z}[u]$ it suffices to have $m \mid \frac{u^{2n}-1}{\sqrt{5}}$ i.e. $m\sqrt{5} \mid u^{2n} - 1$. Now as u is a unit, it is coprime to $m\sqrt{5}$ so theorem 7.4. tells that if $\phi(m\sqrt{5}) \mid 2n$ then $m\sqrt{5} \mid u^{2n} - 1$. Picking such an m, we conclude that $m \mid F_n$ in $\mathbb{Z}[u]$ hence $\frac{F_n}{m} \in \mathbb{Z}[u]$. From the other side, $\frac{F_n}{m} \in \mathbb{Q}$. Therefore $\frac{F_n}{m} \in \mathbb{Q} \bigcap \mathbb{Z}[u] = \mathbb{Z}$, as desired.

Remark. we have used here the fact that for integers a, b, $a \mid b$ in $\mathbb{Z}$ if and only if $a \mid b$ in R where R is some quadratic ring. This is true because of the equality $R \bigcap \mathbb{Q} = \mathbb{Z}$. Indeed, if $u+v\epsilon \in \mathbb{Q}$ for integer u, v then as ϵ is not rational we conclude $v = 0$ so $u + v\epsilon = u \in \mathbb{Z}$.

Example 9.6. (Sankt-Petersburg 239 High-School Olympiad) Consider $(F_n)_{n\in\mathbb{N}}$ be the Fibonacci sequence. Show that if $239 \mid F_n$, then n is even.

Solution. Assume that n is odd. Then $F_n = \frac{u^n+\frac{1}{u^n}}{\sqrt{5}} = \frac{u^{2n}+1}{\sqrt{5}}$ (we keep the notations of the previous example). We thus conclude that $239 \mid u^{2n}+1$ in $\mathbb{Z}[u]$. As $\left(\frac{5}{239}\right) = \left(\frac{239}{5}\right) = 1$, the prime 239 splits in $\mathbb{Z}[u]$. Let π be a prime divisor of 239, of norm 239. Set $z = u^n$, then z is a unit so is coprime to π, hence Fermat's theorem tells that $z^{238} \equiv 1 \pmod{\pi}$. But the condition $239 \mid u^{2n}+1$ implies $z^2 \equiv -1 \pmod{\pi}$ so $z^{238} = (z^2)^{119} \equiv (-1)^{119} = -1 \pmod{\pi}$. Hence we get $1 \equiv -1 \pmod{\pi}$, a contradiction.

Example 9.7. (IMO Shortlist, 2004) Let $m = 4k^2 - 5$. For $a, b \in \mathbb{Z}$ consider the sequence $(x_n)_{n\in\mathbb{N}}$ defined by $x_0 = a, x_1 = b$, $x_{n+2} = x_{n+1} + x_n$. Show that one can choose a, b in such a way that all terms of the sequence are coprime to m.

Solution. Keeping the notations of the previous problems, let's work in $\mathbb{Z}[u]$. Let $p_1, p_2, \ldots, p_n$ be all the distinct primes not dividing m. We see that 5 is a square modulo p_i hence p_i splits into primes π_i and $\overline{\pi_i}$. We need to ensure that none of the terms of the sequence is divisible by any of the π_i (and hence by any of the p_i).

Note that (u^n) satisfies the recurrence, therefore if $a \equiv 1 \pmod{\pi_i}$ and $b \equiv u \pmod{\pi_i}$ then we will have $x_n \equiv u^n \pmod{\pi_i}$ and since u^n is a unit, we will have x_n not divisible by π_i, as desired. Now, from what we have seen in the proof of theorem 6.2., every member of the ring is congruent to some integer modulo π_i, if $N(\pi_i)$ is a prime number (and it is). Therefore one can choose integers k_i such that $k_i \equiv u \pmod{\pi_i}$. By the Chinese Remainder Theorem (in $\mathbb{Z}$), we can find an integer b such that $b \equiv k_i \pmod{p_i}$. This implies $b \equiv u \pmod{\pi_i}$ and so $x_n \equiv u^n \pmod{\pi_i}$, and we are done.

Example 9.8. (Mathematical Reflections, 2007) Solve in integers the equation $x^3 - y^2 = 2$.

Solution. The equation can be rewritten as $x^3 = (y + \sqrt{-2})(y - \sqrt{-2})$. The numbers $y + \sqrt{-2}$ and $y - \sqrt{-2}$ are coprime in $\mathbb{Z}[\sqrt{-2}]$. Indeed, if $\pi \mid y + \sqrt{-2}$ and $\pi \mid y - \sqrt{-2}$ for π prime then $\pi \mid (y - \sqrt{-2}) - (y + \sqrt{-2}) = -2\sqrt{-2} = (\sqrt{-2})^3$ so $\pi = \sqrt{-2}$. Then $\sqrt{-2} \mid y$ meaning that y is even and as $x^3 = y^2 + 2$, x must be even, too. But for x, y even, $x^3 - y^2$ is divisible by 4 while 2 is not. Contradiction.

Since $y + \sqrt{-2}, y - \sqrt{-2}$ are coprime and their product is x^3, each of them must be associated to a cube. In $\mathbb{Z}[\sqrt{-2}]$ the only units are 1 and -1 which are cubes themselves, so $y+\sqrt{-2}$, $y-\sqrt{-2}$ must be cubes in $\mathbb{Z}[\sqrt{-2}]$. If $y+\sqrt{-2} = (a+b\sqrt{-2})^3$ we get $y + \sqrt{-2} = a^3 - 6ab^2 + (3a^2b - 2b^3)\sqrt{-2}$ so $3a^2b - 2b^3 = 1$. As $b \mid 3a^2b - 2b^3$ one deduces $b \mid 1$ so $b = \pm 1$. For $b = 1$ we get $3a^2 - 2 = 1$ so $a = \pm 1$ which lead to the solutions $x = 3, y = 5$ and $x = 3, y = -5$. For $b = 1$ we get $2 - 3a^2 = 1$ so $a^2 = -1$, impossible.

10 Conclusion

As you see, working in quadratic rings may often be invaluably helpful for understanding deep number-theoretical facts about primes and solving interesting mathematical problems. And this was obtained by constructing a theory of divisibility just for quadratic rings, whereas one can add to $\mathbb{Z}$ roots of any polynomial, and even more than one root! It turns out that divisibility theories can be constructed for these rings too, even if the prime decomposition is not unique. This makes the task much more difficult, but also much more fun. I myself find the idea of taming divisibility in algebraic extensions so marvelous, that for me, number theory is the most beautiful part of mathematics (as of now). And I hope you have felt something similar. If you did, and if you want to learn something more, I encourage you to seek and find new sources. But keep in mind that number theory is, in general, a very hard subject, using a lot of other theories for its purposes. Before going into deeper number theory, you should first learn linear algebra and modern algebra. If any of you is interested, send me an e-mail and I can send you an electronic copy of a good book about these subjects. Only after swallowing these two concepts, you should be prepared to go deeper into number theory (at the college level). Again, I can recommend you some books at that level, but you should be well-prepared. Good luck, and I hope these notes were useful or at least interesting to you.