The Number Theory of Binary Quadratic Forms

Nic Ford

Mathcamp 2025

Introduction

These are notes from a class that I taught at Canada/USA Mathcamp 2025. The target audience is a student who is comfortable with mathematical proof and has seen some basic number theory — say enough modular arithmetic to know whether an integer a is invertible modulo another integer n — but who hasn't had any exposure to quadratic forms before. At one point, we'll also use a bit of linear algebra; if you've been exposed to the concept of the determinant of a 2×2 matrix and what it's used for, you should have enough background for this as well. As is often the case with lecture notes, while I've done my best to make them readable in isolation, they might not be as easy to follow outside the context of the class for which they were written.

This class will be all about a single question: given three integers a, b, c, which integers can you write in the form

$$a x^2 + b x y + c y^2$$

for integers x and y? This expression is called a **binary quadratic form**, and as we'll see as the week progresses, there is a lot of deep number theory implicated in such a simple-sounding question. In fact, while we'll be able to make a lot of progress, it's a complex enough problem that we won't be able to give a completely general answer in just one week.

If you're interested in filling in the gaps, I highly recommend looking through a few books on the subject. Some of my favorites that I found while preparing this class were:

- *The Sensual (Quadratic) Form* by John Conway. This is a very slim book that goes through several different ways of looking at a quadratic form. It's a quite elementary and fun read and offers a different perspective than the one we'll be focusing on here.
- *Topology of Numbers* by Allen Hatcher. This one is longer and more detailed, and covers everything we'll be doing in this class and more. It's a nice bridge from Conway's book (which it takes a lot of inspiration from) and the perspective that I take in this class.
- *Primes of the Form* $x^2 + n y^2$ by David Cox. This book, as the title suggests, focuses on only a special case of our central problem. It starts at a fairly elementary level and builds up to some quite advanced material; this might be a nice book to read part of and come back to at a later point in your mathematical education.

1 Sums of Squares

To start, we're going to focus on what's probably the simplest nontrivial special case of this question, one that we will be able to give a complete answer to: which integers can be written as a sum of two squares? (This is the binary quadratic form with $a = c = 1$ and $b = 0$.)

So we're looking for all integers that can be written in the form $x^2 + y^2$ for integers x and y. For any particular small integer, it's not that hard to work out by hand whether this is possible. If you'd like, you can verify that among the integers from 0 to 40, the ones that can written as a sum of two squares are exactly

$$0, 1, 2, 4, 5, 8, 9, 10, 13, 16, 17, 18, 20, 25, 26, 29, 32, 34, 36, 37, 40.$$

One pattern you might notice is that none of these numbers is equivalent to 3 mod 4. Indeed, it's not that hard to verify that no sum of two squares can be 3 mod 4: if you start with a number which is 0, 1, 2, or 3 mod 4 — the only possibilities — and square it, you will always get something that's either 0 or 1 mod 4 as the result, and the sum of two such numbers can only be 0, 1, or 2 mod 4.

This is part of the answer, but it isn't the whole story: we can already see on this list that some numbers that are 0, 1, or 2 mod 4 are missing. For example, 12, 21, and 6 all don't appear. We're going to need a more sophisticated plan of attack.

1.1 The Gaussian Integers

One common strategy for understanding quadratic polynomials is to try to factor them. Our polynomial $x^2 + y^2$ doesn't factor usefully if we restrict ourselves to the real numbers, but if we allow complex numbers, we can write

$$x^2 + y^2 = (x + y\,i)(x - y\,i).$$

These two factors are complex numbers of a special type: complex numbers whose real and imaginary parts are both integers. These are called **Gaussian integers**, and we'll write

$$\mathbb{Z}[i] = \{a + b\,i \in \mathbb{C} : a, b \in \mathbb{Z}\}$$

for the set of all of them.[1]

The question of whether an integer can be written as a sum of two squares is therefore equivalent to the question of whether we can factor it into two Gaussian integers in this way, where the real parts are the same and the imaginary parts are negatives of each other. For example, since $5 = 2^2 + 1^2$, we can write 5 as the product of the Gaussian integers $2 + i$ and $2 - i$. Because of the role they play in the problem we're trying to solve, it will be very helpful to have names for products of this type.

Given a Gaussian integer $z = x + y\,i$, the **conjugate** of z is the Gaussian integer $\overline{z} = x - y\,i$. The **norm** of z is the product

$$N(z) = z\overline{z} = x^2 + y^2.$$

[1] When the distinction between Gaussian integers and ordinary integers is important, you might see the latter called "rational integers" in some number theory texts, but we won't use this terminology. We'll instead always say "integer" for an element of $\mathbb{Z}$ and "Gaussian integer" for an element of $\mathbb{Z}[i]$.

Using this terminology, the question we're trying to answer is whether a given integer occurs as the norm of some Gaussian integer.

In Exercise 1.1 below, you'll verify that $\overline{zw} = \overline{z} \cdot \overline{w}$ for Gaussian integers z, w. This immediately implies an incredibly useful property of norms. We get that

$$N(zw) = zw\overline{zw} = z\overline{z}w\overline{w} = N(z)N(w),$$

that is, the norm of a product of Gaussian integers is the product of the norms. In terms of our sums-of-squares problem, this means:

Proposition 1.1. *If two integers m and n can both be written as a sum of two squares, so can mn.*

1.2 Factorizations

In light of this, it makes sense to turn our attention to the question of how to factor a Gaussian integer. If we understand how a Gaussian integer splits up into factors, then maybe we can try to understand what norms of each of the factors can have and then combine them using our new multiplication rule.

Every ordinary integer can be factored into primes, and these factorizations are always unique up to reordering. We've already seen that a number might factor as a Gaussian integer even if it is prime as an ordinary integer. We saw, for example, that $5 = (2+i)(2-i)$. But we might hope for a similar unique prime factorization result in the Gaussian integers.

There is one fairly trivial way that uniqueness of factorizations can fail in the Gaussian integers. We can, for example, also write $5 = (1+2i)(1-2i)$. But while this looks like a new factorization, if you notice that $1+2i = i(2-i)$ and $1-2i = -i(2+i)$, you'll see that this factorization is actually our old one "in disguise," up to factors of i and $-i$ that end up cancelling.[2]

It will be helpful to have some language to describe this situation. We'll say that a Gaussian integer u is a **unit** if there exists a Gaussian integer v such that $uv = 1$. You'll show in Exercise 1.3 that a Gaussian integer is a unit if and only if it has norm 1. A Gaussian integer p is called **irreducible** if it's not a unit and if, whenever it factors as $p = zw$, one of z or w is always a unit.

The result we're hoping for, then, is that any Gaussian integer factors into irreducibles, and that these factorizations are unique except for reordering the factors and multiplying them by units. This is in fact true, and we'll prove it in the next section in Proposition 2.4. For now, though, let's take it for granted and see how it helps us determine which integers are sums of two squares.

The key step in the argument is a classification of the irreducible Gaussian integers that appear in these factorizations. The answer turns out to be quite nice:

Proposition 1.2. *Let p be a prime integer. There are three possibilities for how p can factor as a Gaussian integer:*

- *We could have $p = \alpha\overline{\alpha}$ for some irreducible Gaussian integer α, where $\overline{\alpha} = u\alpha$ for some unit u. In this case we say p is **ramified**. If instead $\overline{\alpha}$ is* not *a unit multiple of α, we say p is **split**. (Notice that in this case $N(\alpha) = p$.)*
- *Alternatively, p could itself be irreducible as a Gaussian integer. In this case we say p is **inert**. (In this case $N(p) = p^2$.)*

[2] In fact, a similar thing happens with factorizations of ordinary integers: we have both $6 = 2 \cdot 3$ and $6 = (-2)(-3)$. We usually deal with this by just asking for the prime factors to be positive, but this option isn't available to us in the Gaussian integers, hence the need for this business about units.

Furthermore, every irreducible Gaussian integer appears in this way. That is, up to multiplication by units, the complete list of irreducible Gaussian integers is:

- *for every ramified prime p, one irreducible of norm p;*
- *for every split prime p, two irreducibles of norm p which are conjugates of each other; and*
- *for every inert prime p, p itself.*

Proof. Start by factoring our prime integer p into irreducibles, say $p = q_1 q_2 \cdots q_n$. Because (by Exercise 1.3) any Gaussian integer with norm 1 is a unit, all the q_i's have to have norms greater than 1. Taking the norm of both sides, we then see that

$$N(q_1)N(q_2)\cdots N(q_n) = N(p) = p^2.$$

So the $N(q_i)$'s are integers which are bigger than 1 and their product is p^2. Since p is prime, there are only two ways this could happen: there could just be one q_i with norm p^2, or there could be two which each have norm p. In the first case, we just have $p = q_1$, so p is inert. In the second case, $p = q_1 q_2$, where $N(q_1) = N(q_2) = p$. But then by the definition of norm that $p = N(q_1) = q_1\overline{q_1}$, so since $p = q_1 q_2$, we conclude that $q_2 = \overline{q_1}$. We therefore get that p is either ramified or split, depending on whether q_1 and q_2 differ by a unit.

Our last task is to show that the irreducibles that show up in this way account for all the irreducibles in $\mathbb{Z}[i]$. So let q be an irreducible Gaussian integer, and let $n = q\overline{q}$ be its norm. You'll show in Exercise 1.2 that $\overline{q}$ is irreducible as well, so this is the factorization of n into irreducibles. Since q isn't a unit, it's norm is not 1, so there's some prime p that divides n.

But then p in turn is divisible by one of the irreducibles from our list: either p is inert, so it's on our list itself, or it factors into an irreducible and its conjugate. Either way, this irreducible divides $q\overline{q}$, so because factorization is unique, it's either equal to q or $\overline{q}$. Because (as I encourage you to convince yourself) our list of irreducibles is closed under taking conjugates, we can conclude from this that q is on our list. □

1.3 From Factorizations to Sums of Squares

Knowing whether or not a prime p is inert is essentially the same as knowing whether p can be written as a sum of two squares. We saw in the course of the proof that if p is split or ramified then it's the norm of a Gaussian integer, and therefore it's a sum of two squares. Conversely, if $p = N(z)$ for some Gaussian integer z, then $p = z\overline{z}$, and since neither z nor $\overline{z}$ is a unit, p can't be inert. So p is a sum of two squares if and only if it's not inert.

And in fact, we can fairly directly turn it into an answer to the question in general:

Theorem 1.3. *Let m be an integer greater than 1, and factor it into primes as $m = p_1 p_2 \cdots p_n$. Then m is a sum of two squares if and only if each inert prime appears an even number of times in this factorization.*

Proof. Suppose that each inert prime shows up an even number of times. This means we can write our factorization in the form

$$m = q_1 q_2 \cdots q_k r_1^2 r_2^2 \cdots r_l^2,$$

where the q_i's are all either split or ramified and the r_i's are inert. (Some of the factors could be repeated.) From Proposition 1.2, for each q_i there is some irreducible Gaussian integer α_i whose norm is q_i. But then, because the norm is multiplicative,

$$N(\alpha_1\alpha_2\cdots\alpha_k r_1 r_2\cdots r_l) = q_1 q_2\cdots q_k r_1^2 r_2^2\cdots r_l^2 = m.$$

So m is a norm, which is the same as being a sum of two squares.

For the other direction, suppose m is a sum of two squares. This means m is the norm of some Gaussian integer z. We can factor z into irreducibles, and again by Proposition 1.2, we know that the irreducibles that appear all either have norm p for some split or ramified prime p or have norm q^2 for some inert prime q. So, when you multiply the norms of all these irreducibles to find the norm of z, each inert prime will show up an even number of times, which is what we needed to show. □

So where does this leave us? Assuming we can show that the Gaussian integers have unique factorization (which, again, we'll do in Proposition 2.4) Theorem 1.3 gives us a more or less complete answer to the question of which integers can be written as a sum of two squares: it's exactly those numbers whose prime factorizations contain each inert prime an even number of times.

As a concrete answer to the question, it leaves a bit to be desired. It relies on knowing which primes are inert, and this is really just asking which primes are sums of two squares. So it would be fair to say that we'll only *really* be done answering the question when we've found a way to determine whether a given prime is inert. You'll fill in this gap in the exercises below.

Exercises

1.1. Prove that, for Gaussian integers z and w, $\overline{zw} = \overline{z}\cdot\overline{w}$.

1.2. Prove that if q is an irreducible Gaussian integer, then $\overline{q}$ is irreducible as well.

1.3. Use the multiplicativity of the norm to prove that a Gaussian integer is a unit if and only if it has norm 1. Conclude from this that the only units are $1, -1, i$, and $-i$.

1.4. (a) Prove that 2 is ramified.

(b) Prove that, if z is a Gaussian integer and $\overline{z}$ is a unit multiple of z, then z is of the form n, ni, $n+ni$, or $n-ni$ for some integer n. Conclude from this that 2 is the only ramified prime.

1.5. Prove that if p is a prime which is equivalent to 3 mod 4, then p is inert.

1.6. In this problem we're going to show that if p is a prime which is equivalent to 1 mod 4, then p is split. Together with Exercises 1.4 and 1.5, along with Theorem 1.3, this gives a complete answer to the question of which numbers can be written as a sum of two squares.

(a) Prove **Wilson's Theorem**: if p is any prime, then $(p-1)! \equiv -1 \bmod p$. *[Hint: Multiply all the integers from 1 to $p-1$, and try to pair off each one with its inverse mod p. Which ones are left behind, and why?]*

(b) Use Wilson's Theorem to show that, if p is a prime which is equivalent to 1 mod 4, then -1 is a square mod p.

(c) Deduce that, for some integer m, p divides m^2+1. Suppose p is inert. Use unique factorization of Gaussian integers to derive a contradiction.

2 Unique Factorization

Now that we've seen how factoring Gaussian integers into irreducibles can be used to answer our question about sums of two squares, we can turn our attention to the last piece of unfinished business from that story: the proof that Gaussian integers have unique factorization.

If you're used to the integers, you might think that this should be obvious. But this is a less trivial undertaking than it might sound! There are very similar number systems, which we'll explore later on, where unique factorization fails. For example, if we define

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

then in $\mathbb{Z}[\sqrt{-5}]$ we have

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

In Exercise 2.1 you'll show that all four of the numbers on the right side here are irreducible in $\mathbb{Z}[\sqrt{-5}]$. This means 6 factors into irreducibles in two different ways!

So whatever is responsible for unique factorization working out in $\mathbb{Z}[i]$, it will have to involve some feature that differentiates it from $\mathbb{Z}[\sqrt{-5}]$. Our job today will be to explain what that feature is in a way that will set us up to locate it in other number systems that will be useful for analyzing other quadratic forms.

2.1 Division with Remainder

There are lots of ways to prove that the ordinary integers have unique factorization. One of them goes through **division with remainder**, which is just the familiar statement that, if a and m are integers with $m \neq 0$, then there exist integers q and r (the "quotient" and "remainder") such that $a = qm + r$ and $0 \leq r < m$.

Our proof that the Gaussian integers have unique factorization is going to follow the same path.[3] We'll therefore start by showing that something like this division with remainder procedure also works in the Gaussian integers. In the integer version, the remainder has to be smaller than m, so we'll need a similar condition for our Gaussian integer version. One natural choice is to require it to have a smaller *norm*. The complete statement is:

Proposition 2.1. *Let a and m be Gaussian integers, and suppose $m \neq 0$. Then there exist Gaussian integers q and r such that $a = qm + r$ and $N(r) < N(m)$.*

Proof. Since $m \neq 0$, the equation $a = qm + r$ is equivalent to the equation

$$\frac{a}{m} = q + \frac{r}{m}.$$

The requirement that $N(r) < N(m)$ is then the same as requiring $N(r/m) < 1$. If these were ordinary integers, we could pick q and r by letting q be "integer part" of a/m; this would make the difference between a/m and q less than 1, and since this difference is exactly r/m, we'd be done.

So let's let q be the closest Gaussian integer to a/m. (If there's a tie, just pick one.) We'll be done if we can show that we'll always then have $N(r/m) = N(q - a/m) < 1$. This is probably

[3] In fact, if you've never seen a complete proof of unique factorization in the ordinary integers, exactly the same argument will work there too, so you'll get that as an added bonus!

easiest to see geometrically. The Gaussian integers are the vertices of a grid of squares in the complex plane where each square has side length 1. The number a/m will land inside one of these squares, and the distance from any point inside a unit square to its closest vertex is always at most $1/\sqrt{2}$.

Now, if $r/m = x + y\,i$, then the distance in question is $\sqrt{x^2+y^2}$, whereas $N(r/m) = x^2+y^2$. That is, the norm is just the distance squared, so if the distance is at most $1/\sqrt{2}$, the norm is at most $1/2$. This is indeed less than 1, which is all we needed. □

Notice that, unlike the integer version, the q and r we get from this version of division with remainder aren't necessarily unique. For example, taking $a = 7$ and $m = 3$, we could write $7 = 2\cdot 3+1$ or $7 = 3\cdot 3+(-2)$. Both 1 and -2 have norms less than the norm of 3, so they're both suitable remainders.

2.2 Bézout's Lemma and Unique Factorization

The reason division with remainder is helpful for proving unique factorization in the ordinary integers is that it gives a simple proof of **Bézout's Lemma**, the statement that, if a and b are two integers with no common factors, then it's possible to find integers m and n such that $am + bn = 1$. The same thing is true in the Gaussian integers, and for the same reason:

Lemma 2.2. *Let a and b be Gaussian integers, and suppose a and b have no common factors other than units. Then there exist Gaussian integers m and n such that $am + bn = 1$.*

Proof. For some choice of a and b, consider the set S of all Gaussian integers of the form $am + bn$. Since the norms of the elements of S are all nonnegative integers, there has to be an element d with the smallest nonzero norm of all the elements of S. Since $d \in S$, it can be written in the form $am + bn$ for some Gaussian integers m and n.

I claim that d is a common factor of a and b. To see this, let's see what happens when we apply the division-with-remainder procedure to a and d: we get Gaussian integers q and r with $a = dq + r$ and $N(r) < N(d)$. But since $d = am + bn$, a bit of algebra shows us that $r = a - dq = a(1-qm) + b(qn)$. This means r is an element of S with smaller norm than d, so since d had the smallest possible nonzero norm, we must have $N(r) = 0$. But 0 is the only Gaussian integer with norm 0, so in fact $r = 0$, meaning $a = dq$.

Exactly the same argument will show that b is also a multiple of d. Since the only common factors of a and b are units, this means d is a unit, so suppose $de = 1$. Multiplying the equation $d = am + bn$ by e then gives $1 = a(em) + b(en)$, giving us our desired expression for 1 as the sum of a multiple of a and a multiple of b. □

Bézout's Lemma leads more or less directly to an important fact about irreducibles which will be the main ingredient in our proof of unique factorization:

Corollary 2.3. *Suppose p is an irreducible Gaussian integer, and, for some Gaussian integers a and b, ab is a multiple of p. Then either a or b is a multiple of p.*

Proof. Suppose ab is a multiple of p, but a isn't. We'll show that b is a multiple of p.

Because p is irreducible, the only non-unit factor of p (up to multiplying by a unit) is p itself, which by assumption isn't a factor of a, so a and p have no common factors other than units. We can therefore apply Lemma 2.2 to find Gaussian integers m and n with $am + pn = 1$.

But if we multiply this equation by b, we see that $b = abm + pbn$. Both factors on the right are multiples of p: the second obviously so, and the first because we assumed that ab is a multiple of p. Therefore b is a multiple of p as well. □

In a number system like this, a distinction is often made between **irreducibles** and **primes**. An irreducible is a number which can't be factored into non-units, whereas a prime is a number with the property in the statement of Corollary 2.3. The example from the beginning of the section shows that, in $\mathbb{Z}[\sqrt{-5}]$, 2 is irreducible but not prime.

We now have everything we need to prove unique factorization.

Proposition 2.4. *The Gaussian integers have unique factorization into irreducibles. That is, for any nonzero Gaussian integer z that isn't a unit, we can write $z = p_1 p_2 \cdots p_n$ for some irreducibles p_i, and the irreducibles that appear in this way are unique up to reordering and multiplying by units.*

Proof. We'll first show that factorizations exist at all. If z is irreducible, then we're done: $z = z$ is already a factorization of z into irreducibles. Otherwise, it's possible to write $z = w w'$ where neither w nor w' is a unit. We know that $N(z) = N(w)N(w')$, and by Exercise 1.3 from the last section, since neither factor is a unit, that their norms are both bigger than 1. Since their product is $N(z)$, this means $N(w)$ and $N(w')$ are both strictly smaller than $N(z)$.

Continuing in this way, if w and w' are irreducible, then we're done, otherwise take whichever one isn't irreducible and factor it. Because the norms of the factors are positive integers and they keep decreasing every time a factor fails to be irreducible, this process has to end after finitely many steps, and when this happens, we have our factorization of z into irreducibles.[4]

Now suppose we had two different factorizations of z into irreducibles, say

$$z = p_1 p_2 \cdots p_n = q_1 q_2 \cdots q_m.$$

Then p_1 is a factor of $q_1 \cdot (q_2 \cdots q_m)$, so by Corollary 2.3, it's either a factor of q_1 or it's a factor of $q_2 \cdots q_m$. If we're in the second case, we can apply the same logic again to conclude that p_1 is either a factor of q_2 or of $q_3 \cdots q_m$. Continuing in this way, we can conclude that one of the q_i's has to be a multiple of p_1.

Suppose q_i is a multiple of p_1. Since both p_1 and q_i are irreducible, this just means they're unit multiples of each other. Therefore, after possibly multiplying by a unit, we have the same factor showing up in both of our factorizations, and so we can cancel it from the equation, getting that $p_2 \cdots p_n = q_1 \cdots q_{i-1} q_{i+1} \cdots q_m$. Now, the same argument shows that one of the remaining q_j's is (up to multiplication by a unit) equal to p_2, so we can cancel p_2 and q_j from the equation. Every time we do this, the two factorizations get shorter, and in the end we get that every one of the p's is a unit multiple of a different one of the q's.[5] Our two factorizations are therefore the same up to multiplying by units and reordering, which was exactly what we needed to show. □

This finishes our proof that Gaussian integers have unique factorization. In the next sections, we're going to see how to generalize the ideas that went into the sums-of-squares proof to other binary quadratic forms, and we'll see that unique factorization plays a vital role there as well.

Exercises

2.1. Prove that 2, 3, $1+\sqrt{-5}$, and $1-\sqrt{-5}$ are irreducibles in $\mathbb{Z}[\sqrt{-5}]$, and that none of these is a unit multiple of any other. (Once we know this, the fact that $2 \cdot 3 = (1+\sqrt{-5})(1-\sqrt{-5})$ really does show that unique factorization fails here.)

[4] Notice that this proof of the existence of factorizations didn't actually require anything we proved in this section! It'll be in the proof of uniqueness that they come up.

[5] Why does this argument guarantee that we can't assign two different p's to the same q?

2.2. Suppose that, instead of working in the Gaussian integers, we worked in

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

for some positive integer d. If we tried to prove a version of our division with remainder result (Proposition 2.1) using the same argument, for which values of d would we succeed?

2.3. (a) Use Bézout's Lemma to prove that Gaussian integers have gcd's. That is, given two nonzero Gaussian integers a and b, prove that there exists a Gaussian integer d which (i) is a common divisor of a and b, and (ii) has the property that any other common divisor of a and b also divides d.

(b) Prove that gcd's are unique up to multiplication by units.

(c) Prove that the set $\{am + bn : m, n \in \mathbb{Z}[i]\}$ is exactly the set of multiples of d.

2.4. Our proof of unique factorization required using Corollary 2.3. Prove that the implication in the other direction is true as well, that is, if you assume unique factorization holds, then Corollary 2.3 follows.

3 Other Quadratic Forms

3.1 Equivalent Forms

An arbitrary binary quadratic form looks like $a\,x^2+b\,x\,y+c\,y^2$ for some integers a, b, c. Given such a form, our goal will be to figure out which integers it **represents**, that is, which integers can be written as $a\,x^2+b\,x\,y+c\,y^2$ for integers x and y.

There's a relatively straightforward way to turn one of these forms into another one which represents exactly the same set of numbers. Suppose, for example, we start with the form x^2+y^2 and do a linear change of coordinates like $x'=x+y$, $y'=y$. If we plug $x=x'-y'$ and $y=y'$ into our form x^2+y^2, we get the form $x'^2-2x'y'+2y'^2$.

If you had just seen this final expression, you might not have guessed that the integers it represents are exactly the sums of two squares, but the method we used to construct it makes this clear: given any number of the form $n=x'^2-2x'y'+2y'^2$, we can find the corresponding pair of integers $(x,y)=(x'-y',y')$ and conclude that $n=x^2+y^2$.

When will this procedure work in general? If we try to use a linear change of coordinates $(x',y')=(r\,x+s\,y,t\,x+u\,y)$ for some integers r,s,t,u, we need to have an inverse change of coordinates, and because we're specifically interested in plugging integers into our quadratic forms, we need the inverse coordinate change to also have integer coefficients.

There's a simple way to tell when this will happen that uses the formula for the inverse of a 2×2 matrix: the inverse of the matrix $\left(\begin{smallmatrix} r & s \\ t & u \end{smallmatrix}\right)$ is $\frac{1}{r\,u-s\,t}\left(\begin{smallmatrix} u & -s \\ -t & r \end{smallmatrix}\right)$. So in order to guarantee that the inverse coordinate change has integer coefficients, we want $r\,u-s\,t=\pm 1$.

If two forms are related by a linear change of coordinates in this way, with $r\,u-s\,t=\pm 1$, then we'll say the two forms are **equivalent**. (For example, our discussion above shows that the forms x^2+y^2 and $x^2-2x\,y+2y^2$ are equivalent.) Because both the initial change of coordinates and its inverse have integer coefficients, they will both take integers to integers, and this means that two equivalent forms will always represent the same set of numbers.

This concept will be a very useful tool for us; a lot of facts about which numbers can be represented by a given form will be easier to see after replacing it with an equivalent form.

Many of those results involve representations of a particular type: we'll say that an integer n is **primitively represented** by our quadratic form if $n=a\,x^2+b\,x\,y+c\,y^2$ for some x and y with no common factors. Much of the reason this definition is worth thinking about comes from the following result.

Proposition 3.1. *Suppose n is primitively represented by the quadratic form $a\,x^2+b\,x\,y+c\,y^2$. Then this form is equivalent to a form whose first coefficient is n.*

Proof. Say $n=a\,w^2+b\,w\,z+c\,z^2$ where w and z have no common factors. Bézout's Lemma then lets us find integers s,u with $w\,s+z\,u=1$. This means that the coordinate change $(x,y)=(w\,x'-u\,y',z\,x'+s\,y')$ has determinant 1, that is, it's of the type that will produce an equivalent form.

That equivalent form will look like $a'x'^2+b'x'y'+c'y'^2$ for some integers a',b',c'. If you plug $x'=1$ and $y'=0$ into this form, you get a', the first coefficient. But the expression for our coordinate change implies that this corresponds to $(x,y)=(w,z)$, which were the inputs that produced n. So $a'=n$. □

3.2 The Discriminant

A natural question to ask is whether there's an easy way to tell whether or not two forms are equivalent. The following fact is a big step toward answering that question.

Proposition 3.2. *The **discriminant** of the form $ax^2+bxy+cy^2$ is the number $\Delta = b^2-4ac$. Two equivalent forms always have the same discriminant.*

Proof. It's possible to prove this using a direct computation, but the proof is much nicer if we use a bit of linear algebra. Plugging a pair of numbers into our quadratic form can be written as a matrix multiplication: I encourage you to verify that

$$ax^2+bxy+cy^2 = \begin{pmatrix} x & y \end{pmatrix} A \begin{pmatrix} x \\ y \end{pmatrix},$$

where $A = \begin{pmatrix} a & \frac{b}{2} \\ \frac{b}{2} & c \end{pmatrix}$. When written this way, $\Delta = -4\det A$, so it will be enough to show that our coordinate change leaves this determinant unchanged.

Suppose $ax^2+bxy+cy^2$ is equivalent to $a'x'^2+b'x'y'+c'y'^2$ via the coordinate change $(x,y) = (rx'+sy', tx'+uy')$, so that $\begin{pmatrix} x \\ y \end{pmatrix} = C\begin{pmatrix} x' \\ y' \end{pmatrix}$, where $C = \begin{pmatrix} r & s \\ t & u \end{pmatrix}$. Recall that, by our definition of equivalent forms, $\det C = \pm 1$. The row vectors are then related in a similar way, but with the transpose: $\begin{pmatrix} x & y \end{pmatrix} = \begin{pmatrix} x' & y' \end{pmatrix} C^T$.

If we plug in both of these expressions, we get that

$$a'x'^2+b'x'y'+c'y'^2 = \begin{pmatrix} x' & y' \end{pmatrix} C^T A C \begin{pmatrix} x' \\ y' \end{pmatrix}.$$

Since taking the transpose doesn't change the determinant, $\det(C^TAC) = (\det C)^2 \det A$, and since $\det C = \pm 1$, this is just $\det A$, which is what we needed to show. □

This result tells us that, if two forms are equivalent, the discriminants have to be the same. But it doesn't imply that two forms with the same discriminant have to be equivalent, and in fact, this is not always true. For example, the forms x^2+4y^2 and $2x^2+2y^2$ both have discriminant -16, but the first one represents 1 and the second doesn't, since all its outputs are even.

In our discussion of the quadratic form x^2+y^2, we were able to reduce everything to the question of which *primes* could be represented. In particular, we showed that an integer was representable if and only if it was a product of representable primes times a square. It would be nice if a similar thing was true of other quadratic forms, but here the news is once again not good. The form x^2+5y^2 represents 6, so if things were going to work out like they did for sums of squares, we'd expect it to also represent 2 and 3, but I encourage you to prove to yourself that it doesn't.

So, at least for the time being, we're going to have to restrain our ambitions. While the project of reducing everything to primes won't work out for every quadratic form, there is a correspondingly nice story if we just ask which numbers can be represented by *some* form with a given discriminant. We'll say that a number n is **represented in discriminant** Δ if there exists a quadratic form with discriminant Δ representing n.

Lots of our sums-of-squares results have analogues in this new context. You'll fill in most of this picture in the exercises, but here are a couple of results to give you a taste of what's involved.

Proposition 3.3. *Suppose n is primitively represented in discriminant Δ. Then every factor of n is also primitively represented in discriminant Δ.*

Proof. Suppose n is primitively represented by the form $a x^2 + b x y + c y^2$ of discriminant Δ. By Proposition 3.1, there is an equivalent quadratic form which looks like $n x^2 + b' x y + c' y^2$ for some b' and c'. By Proposition 3.2, this new form also has discriminant Δ.

Now, suppose d is a factor of n, say $n = d e$. We can "move the e over" to the last term, replacing $d e x^2 + b' x y + c' y^2$ with $d x^2 + b' x y + e c' y^2$, and this won't change the discriminant. But this form clearly primitively represents d by setting $x = 1$ and $y = 0$. □

Proposition 3.4. *A number n is primitively represented in discriminant Δ if and only if Δ is a square mod $4n$.*

Proof. Suppose Δ is a square mod $4n$, say $\Delta \equiv b^2 \bmod 4n$. This means that, for some integer c, $b^2 - \Delta = 4nc$, or rearranging $\Delta = b^2 - 4nc$. So Δ is the discriminant of the quadratic form $n x^2 + b x y + c y^2$, which clearly primitively represents n.

Conversely, if n is primitively represented by a form of discriminant Δ, we can, like in the proof of the previous result, find an equivalent form with the same discriminant whose first coefficient is n. This form looks like $n x^2 + b x y + c y^2$ for some integers b, c, and since its discriminant is Δ, we see that $\Delta = b^2 - 4nc$, so $\Delta \equiv b^2 \bmod 4n$. □

The final story, which you'll prove in Exercise 3.4, is that a number n is represented in discriminant Δ if and only if, in its prime factorization, each prime which *can't* be represented appears an even number of times. This is very similar to what we proved about sums of squares: in that case, the non-representable primes were the ones that are 3 mod 4.[6]

The difference is that that earlier result was about one specific form, namely $x^2 + y^2$, whereas the results we're proving here only tell you whether a number is represented by *some* form of discriminant Δ. Accomplishing our goal of knowing which numbers are represented by *one specific* quadratic form will require some more work, and in fact we won't be able to give a completely general answer in this class. There are, however, some special cases where there is a very nice story to tell, and that is the task we'll take up next.

Exercises

3.1. Prove that an integer Δ is the discriminant of some quadratic form if and only if Δ is 0 or 1 mod 4.

3.2. (a) Suppose n is square-free, that is, none of the prime factors of n appears more than once. Prove that if n is represented by some quadratic form, then it is primitively represented by that quadratic form.

(b) Suppose n is represented by some quadratic form. Prove that we can write $n = d^2 m$ for some integers d and m, where m is primitively represented by that quadratic form.

(c) Prove that if m is represented by some quadratic form and d is any integer, then $d^2 m$ is also represented by that quadratic form.

3.3. For parts (a) and (c) of this exercise, it will be helpful to use the Chinese Remainder Theorem. If you've never seen it, or you'd just like a reminder of how it works, come find me!

[6] In fact, if $\Delta = -4$, Proposition 3.4 implies that a prime p is representable if and only if -4 is a square mod $4p$, and if p is odd, this is the same as -1 being a square mod p, which was exactly the condition you used in Exercise 1.6 to show that every prime which is 1 mod 4 is represented by $x^2 + y^2$.

(a) Suppose n is odd. Prove that n is primitively represented in discriminant Δ if and only if Δ is a square mod n.

(b) If $\Delta = -4$, as it is for the form $x^2 + y^2$, then for $n = 9$ this condition would require checking whether -4 is a square mod 9, i.e., whether 5 is a square mod 9. The squares mod 9 are just 0, 1, 4, and 7, but 9 is a sum of two squares, namely $3^2 + 0^2$. What's going on here?

(c) Suppose d and e are both primitively represented in discriminant Δ and have no common factors. Prove that de is primitively represented in discriminant Δ.

3.4. Use the results from this section and the previous exercises to prove that an integer n can be represented in discriminant Δ if and only if, in the prime factorization of n, each prime p for which Δ is not a square mod $4p$ appears an even number of times.

3.5. In this section, we mentioned that the quadratic form $x^2 + 5y^2$ represents 6, but not either of its prime factors 2 or 3. This form has discriminant -20. Use the procedures in Propositions 3.1 and 3.3 to find two forms of discriminant -20, one that represents 2 and one that represents 3.

4 Quadratic Number Rings

4.1 Factoring a Quadratic Form

We were led to look at the Gaussian integers by factoring the expression x^2+y^2 as $(x+yi)(x-yi)$. So let's take an arbitrary quadratic form $ax^2+bxy+cy^2$ and try to factor it in a similar way.

If we first factor out y^2, then this expression looks like

$$y^2\left(a\left(\frac{x}{y}\right)^2+b\frac{x}{y}+c\right).$$

The bit in parentheses is just the quadratic polynomial ax^2+bx+c with x/y plugged in, so we can factor it with the quadratic formula. Writing $\Delta=b^2-4ac$ as before, we get

$$ax^2+bxy+cy^2=a\left(x+\frac{b+\sqrt{\Delta}}{2a}y\right)\left(x+\frac{b-\sqrt{\Delta}}{2a}y\right).$$

To make our comparison with the Gaussian integers more direct, we can pull the a's out of the denominators and write it like this:

$$\frac{1}{a}\left(ax+\frac{b+\sqrt{\Delta}}{2}y\right)\left(ax+\frac{b-\sqrt{\Delta}}{2}y\right).$$

If Δ is a square, then these factors will actually be rational numbers. The case where this doesn't happen is much more interesting, so from now on we'll always assume that Δ isn't a square.

Just as the Gaussian integers deserved our attention because they formed the set of numbers that could appear as one of the factors in the factorization of x^2+y^2, we want a set of numbers to serve the same role for this new factorization.

To get the exact set we want, it will be helpful to separate the cases where Δ is even or odd. In fact, because $\Delta=b^2-4ac$ and squares can only be 0 or 1 mod 4, Δ can also only be 0 or 1 mod 4.

If Δ is 0 mod 4, then the formula $\Delta=b^2-4ac$ indicates that b is also even. I encourage you to verify that this means both of the factors in our factorization can be written in the form $r+s\frac{\sqrt{\Delta}}{2}$ where r and s are integers.

If Δ is 1 mod 4, then the same logic implies that b has to be odd. In this case, I encourage you to check that our factors are both of the form $r+s\frac{1+\sqrt{\Delta}}{2}$ where r and s are integers.

So, if we define

$$\tau_\Delta=\begin{cases}\frac{\sqrt{\Delta}}{2} & \text{if } \Delta\equiv 0 \bmod 4\\ \frac{1+\sqrt{\Delta}}{2} & \text{if } \Delta\equiv 1 \bmod 4,\end{cases}$$

then our two factors will always be in the set

$$R_\Delta=\{r+s\tau_\Delta : r,s\in\mathbb{Z}\}.$$

We'll call R_Δ the **quadratic number ring** with discriminant Δ. Here are a couple examples:

- If we start with the quadratic form x^2+y^2, we're supposed to get the Gaussian integers. And indeed, we get $\Delta=-4$, so $\tau_\Delta=\frac{\sqrt{-4}}{2}=i$ and $R_\Delta=\{x+yi : x,y\in\mathbb{Z}\}$ just as expected.

- Similarly, the quadratic form $x^2 - ny^2$ gives us $\Delta = 4n$ and so $\tau_\Delta = \sqrt{n}$. In this case, $R_\Delta = \{r + s\sqrt{n} : r, s \in \mathbb{Z}\}$. Indeed, this is what we would have expected from factoring $x^2 - ny^2 = (x + y\sqrt{n})(x - y\sqrt{n})$.

- Starting with $x^2 + xy + y^2$, we get $\Delta = -3$ and $\tau_\Delta = \frac{1+\sqrt{-3}}{2}$. This point is the vertex of an equilateral triangle whose other vertices are at 0 and 1, so the points in R_Δ form a triangular lattice in the plane. These are called the **Eisenstein integers**.

4.2 Norms and Principal Forms

The arithmetic of R_Δ has a lot in common with the Gaussian integers. R_Δ always contains the integers, and if you add or multiply two elements of R_Δ — that is, two numbers of the form $r + s\tau_\Delta$ — you get another element of R_Δ.[7]

The Gaussian integers were useful because the integers of the form $x^2 + y^2$ were exactly the norms of Gaussian integers. So we'd like a conjugation operation and a norm function on R_Δ that can serve the same role. If we look back at our factorization

$$ax^2 + bxy + cy^2 = \frac{1}{a}\left(ax + \frac{b + \sqrt{\Delta}}{2}y\right)\left(ax + \frac{b - \sqrt{\Delta}}{2}y\right),$$

we can see what we need: conjugation ought to negate the coefficient on $\sqrt{\Delta}$, since that's how to turn the first factor into the second. If Δ is even, this is the same as setting $\overline{\tau_\Delta} = -\tau_\Delta$, and if Δ is odd we end up with $\overline{\tau_\Delta} = \overline{(1 + \sqrt{\Delta})/2} = (1 - \sqrt{\Delta})/2 = 1 - \tau_\Delta$.

If we then define $N(z) = z\overline{z}$, we can write

$$ax^2 + bxy + cy^2 = \frac{1}{a}N\left(ax + \frac{b + \sqrt{\Delta}}{2}y\right).$$

This is a more complicated relationship than we had before. Whereas the integers of the form $x^2 + y^2$ were exactly the norms of Gaussian integers, we can't always conclude from this that the integers of the form $ax^2 + bxy + cy^2$ are exactly the norms of elements of R_Δ. There are in fact two obstacles: the $1/a$ multiplying the norm in this expression, and the fact that there might be elements of R_Δ that can't be written in the form $ax + \frac{b+\sqrt{\Delta}}{2}y$.

Indeed, our study of equivalent quadratic forms in the last section already told us that it's possible for two forms to have the same discriminant but not represent the same set of numbers. If a quadratic form *does* happen to have the property that the numbers it represents are exactly the norms of elements of R_Δ, like we saw for $x^2 + y^2$, we'll call it a **principal form**.

Even without the technology from the last section, it's not hard to see that not every quadratic form is principal. If a form is principal, it will always take the value 1 for some choice of inputs. This is simply because $N(1) = 1$, so 1 is always a norm. This means any form that doesn't represent 1 can't be principal; for example, as I encourage you to verify, this includes $ax^2 + cy^2$ whenever both a and c are strictly greater than 1.

[7] Proving this for addition is pretty transparent; for multiplication, the odd-discriminant case is slightly less obvious. I encourage you to verify it!

4.3 The Role of Unique Factorization

In our story about sums of squares, the key to reducing everything to primes was that the Gaussian integers had unique factorization. This step was crucial in our proof of Proposition 1.2, where we formed an exhaustive list of the irreducibles in $\mathbb{Z}[i]$.

The next big result that we'll prove is that, if R_Δ is a similarly nice ring, then we can expect the same behavior from any quadratic form of discriminant Δ. (This theorem is in fact most of the reason it's worth introducing R_Δ to our discussion in the first place!) It will turn out to be easier to phrase the theorem in terms of Bézout's Lemma rather than unique factorization directly.

Theorem 4.1. *Suppose R_Δ satisfies Bézout's Lemma, that is, for any $a, b \in R_\Delta$ with no common factors, it's possible to find $m, n \in R_\Delta$ such that $am + bn = 1$. Then every quadratic form with discriminant Δ is either a principal form or the negative of a principal form.*

Proof. Consider a quadratic form $ax^2 + bxy + cy^2$ of discriminant Δ. We want to show that the set of numbers of the form

$$\frac{1}{a} N\left(ax + \frac{b+\sqrt{\Delta}}{2} y\right)$$

for integers x, y is exactly the set of norms of elements of R_Δ (or their negatives).

We'll start by looking at which elements of R_Δ can appear inside the norm in this expression. Let's call this set

$$J = \{ax + \frac{b+\sqrt{\Delta}}{2} y : x, y \in \mathbb{Z}\}.$$

In Exercise 4.1, you'll show that J can also be written as $\{ax + \frac{b+\sqrt{\Delta}}{2} y : x, y \in R_\Delta\}$, that is, we can allow x and y to be arbitrary elements of R_Δ rather than just integers and the resulting set will stay the same.

You showed in Exercise 2.3 that Bézout's Lemma implies the existence of gcd's.[8] So let d be the gcd of a and $\frac{b+\sqrt{\Delta}}{2}$. Since we now know that we can write $J = \{ax + \frac{b+\sqrt{\Delta}}{2} y : x, y \in R_\Delta\}$, that exercise tells us that J consists of exactly the multiples of d.

We claim that in fact $N(d) = \pm a$. To see this, first note that $N(d) = d\overline{d}$. Since this is a multiple of d, it's in J, so we can write $N(d) = am + \frac{b+\sqrt{\Delta}}{2} n$ for *integers* m, n. And in fact, since $N(d)$ is an integer, we can see that $n = 0$, so $N(d) = am$. We also know that, since $a \in J$, it's a multiple of d, say $a = zd$ for some $z \in R_\Delta$.

Taking the conjugate of this equation (and using the fact that a, being an integer, is its own conjugate) we get $a = \overline{z}\overline{d}$. Combining this with our previous result gives $d\overline{d} = am = \overline{z}\overline{d}m$, and so $d = \overline{z}m$. This means that m divides d, and therefore m divides every element of J.

But one of the elements of J is $\frac{b+\sqrt{\Delta}}{2}$, and this is supposed to be a multiple of the integer m. If we write this element in terms of τ_Δ, the coefficient on τ_Δ will be 1. So the only integers that could possibly divide it are 1 and -1, so $m = \pm 1$. We then conclude that $N(d) = am = \pm a$, as desired.

This in fact basically finishes the entire proof. Since we know that J consists of exactly the multiples of d, the set of values of our quadratic form consists of all numbers of the form $\frac{1}{a} N(wd)$ for $w \in R_\Delta$. If $N(d) = a$, then this is just $N(w)$, so our form represents exactly the norms of elements of R_Δ. If $N(d) = -a$, then we get $-N(w)$, and our form represents the negatives of the norms of elements of R_Δ. These were the two possibilities laid out in the statement of the theorem. □

[8] That exercise of course was specifically about the Gaussian integers, but the proof only required knowing that Bézout's Lemma was true, so it will work for us as well.

The fact that this theorem can't tell the difference between the form being principal or its negative being principal is an inconvenience, but there are a couple situations where it doesn't matter. If there is an element $\eta \in R_\Delta$ with norm -1, then the set of norms and the set of negatives of norms are the same set, because for any z, $N(\eta z) = -N(z)$. Alternatively, you'll show in Exercise 4.2 that if $\Delta < 0$, then all norms in R_Δ are nonnegative, which makes the two cases trivial to tell apart.

If we're in either of these situations and we also have Bézout's Lemma, then, combined with the results from the last section, we can get a complete answer to our central question: every quadratic represents the same set of numbers, so determining which numbers are represented in discriminant Δ is the same as determining which numbers are represented by one particular form of discriminant Δ. For example, you'll prove in Exercise 4.3 that the ring R_{-3} of Eisenstein integers is one of these, so we can give a complete classification of the integers of the form $x^2 + xy + y^2$.

This of course leaves open the question of what happens when R_Δ is not so nice. We already have an example in the ring $R_{-20} = \mathbb{Z}[\sqrt{-5}]$, where 6 factors as both $2 \cdot 3$ and $(1+\sqrt{-5})(1-\sqrt{-5})$, and we have also correspondingly seen that the form $x^2 + 5y^2$ of discriminant -20 represents 6 but doesn't represent either 2 or 3.

In the final section, we'll take up this specific example in more detail and see how the failure of unique factorization makes the classification of integers of the form $x^2 + 5y^2$ more complicated.

Exercises

4.1. Suppose $ax^2 + bxy + cy^2$ is a quadratic form of discriminant Δ, and, as in the proof of Theorem 4.1, define $J = \{ax + \frac{b+\sqrt{\Delta}}{2}y : x, y \in \mathbb{Z}\}$. Prove that $a\tau_\Delta$ and $\frac{b+\sqrt{\Delta}}{2}\tau_\Delta$ are both in J, and conclude from this that we could also write $J = \{ax + \frac{b+\sqrt{\Delta}}{2}y : x, y \in R_\Delta\}$.

4.2. Prove that, if $\Delta < 0$, then the norm of every element of R_Δ is nonnegative. Conclude that every quadratic form of discriminant Δ either only represents nonnegative numbers or only represents nonpositive numbers.

4.3. In this problem, we'll see what the tools we've developed tell us about integers of the form $x^2 + xy + y^2$.

(a) The discriminant of this form is -3. Argue that the ring R_{-3} has a division-with-remainder theorem and therefore that it satisfies Bézout's Lemma. *[Hint: Proposition 2.1 used the fact that the Gaussian integers formed a square grid in the complex plane. For this one, use the fact that the points of R_{-3} form a triangular grid.]*

(b) What condition on a prime p will tell you whether it's represented by the form $x^2 + xy + y^2$? Check this condition against the first few primes and, for the ones that are representable, find a representation.

(c) How could you determine which integers are represented by the form $x^2 + xy + y^2$?

4.4. In this problem, we'll show that if $\Delta \equiv 0 \bmod 4$ and $\Delta < -8$, then R_Δ never has unique factorization.

(a) Say $\Delta = 4d$ with $d < -2$, so that $R_\Delta = \mathbb{Z}[\sqrt{d}]$. Prove that 2 is not a norm of any element of $\mathbb{Z}[\sqrt{d}]$, and conclude from this that 2 is irreducible.

(b) Show, by factoring $d^2 - d$ in two different ways, that $\mathbb{Z}[\sqrt{d}]$ doesn't have unique factorization.

5 A Taste of the General Case

The story we've just finished telling is a quite compelling one when it works out. If R_Δ satisfies Bézout's Lemma, then every quadratic form of discriminant Δ either represents the norms of elements of R_Δ or their negatives. And, if $\Delta < 0$ or there's an element of norm -1, we can combine this with our earlier results about which numbers are represented in discriminant Δ to get a complete answer to our question of which numbers are represented by a given form.

But R_Δ doesn't always satisfy Bézout's Lemma. Indeed, we've already looked a bit at the example of x^2+5y^2, where $\Delta = -20$, and seen that this complicates our story. In this final section, we're going to give a quick overview — without proving much — of how the algebra of R_Δ can be used to answer the question of which integers can be represented by a given quadratic form of discriminant Δ.

This section will necessarily be much sketchier than what came before it! If you're interested in filling in the details, I highly recommend checking out the books mentioned in the introduction; the last chapter of Hatcher in particular is a good source for this material.

5.1 An Example

Let's start by investigating x^2+5y^2 the same way we started looking at sums of squares, by listing some of the integers that it represents. The numbers represented by this form up to 200 are:

$$0, 1, 4, 5, 6, 9, 14, 16, 20, 21, 24, 25, 29, 30, 36, 41, 45, 46, 49, 54, 56, 61, 64, 69, 70, 80, 81, 84, 86, 89, 94,$$
$$96, 100, 101, 105, 109, 116, 120, 121, 125, 126, 129, 134, 141, 144, 145, 149, 150, 161, 164, 166, 169,$$
$$174, 180, 181, 184, 189, 196.$$

We know that, if a form represents some number, then it also represents any perfect square times that number. So let's trim this list down by removing everything that's a perfect square multiple of another entry on the list. We're left with:

$$1, 5, 6, 14, 21, 29, 30, 41, 46, 61, 69, 70, 86, 89, 94, 101,$$
$$105, 109, 129, 134, 141, 145, 149, 161, 166, 174, 181.$$

If we restrict just to the primes, we get

$$5, 29, 41, 61, 89, 101, 109, 149, 181.$$

These are exactly the primes up to 200 that are 1 or 9 mod 20, along with 5, so we can maybe conjecture that that's how to determine which primes are represented by our form. But, unlike in the case where R_Δ satisfies Bézout's Lemma, this is not enough to determine which *numbers* are represented by our form! Our original list contains numbers, like 6 and 14, whose prime factors aren't on the list of primes we just produced.

If we list the prime factors that *do* show up[9], we get, along with the list above, the primes

$$2, 3, 7, 23, 43, 47, 67, 83, 103, 107, 127, 163, 167.$$

[9] Producing this list requires looking at numbers represented by our form that are bigger than 200. For example, even though 167 isn't a factor of any of the numbers we listed earlier, it is a factor of $334 = 17^2 + 5 \cdot 3^2$.

These are the primes that are 3 or 7 mod 20, along with 2. A bit more playing around will eventually lead to the observation that, in order for a number to be represented by x^2+5y^2, it needs to contain an *even number* of primes from this second list. For example, $129 = 3\cdot 43$, $161 = 7\cdot 23$, and $966 = 2\cdot 3\cdot 7\cdot 23 = 31^2+5\cdot 1^2$ all pass this test, but $42 = 2\cdot 3\cdot 7$ doesn't.

In fact, this second list contains exactly the primes represented by $2x^2+2xy+3y^2$. The elements up to 200 represented by this form are

$$0,2,3,7,8,10,12,15,18,23,27,28,32,35,40,42,43,47,48,50,58,60,63,67,72,75,82,83,87,90,92,$$
$$98,103,107,108,112,115,122,123,127,128,135,138,140,147,160,162,163,167,168,172,175,$$
$$178,183,188,192,200.$$

Let's write S_1 for the set of numbers represented by x^2+5y^2 and S_2 for those represented by $2x^2+2xy+3y^2$. A bit more experimentation would lead you to the conclusion that, if you multiply two numbers from S_2, the result is always in S_1. We can in fact form a "multiplication table" which tells us the result of all such multiplications: multiplying two S_1's or two S_2's always gives an S_1, and multiplying one of each gives an S_2.

If we could establish this, and we could establish which primes belong to each list, we'd have a complete classification of which numbers are represented by x^2+5y^2: n is represented by this form if and only if, in its prime factorization, each prime that isn't represented in discriminant -20 appears an even number of times, and we have an even number of (possibly different) primes from S_2, with no restriction in the primes from S_1.

5.2 Ideals

Let's now turn to the question of what this story should look like in general. Start by once again picking a quadratic form, and since we might be considering more than one form at the same time, let's give this one a name, say

$$Q(x,y) = ax^2+bxy+cy^2.$$

Suppose Q has discriminant Δ. To make everything simpler, let's assume that $\Delta<0$ and that Q represents only nonnegative numbers. In particular, I encourage you to verify that this implies that both a and c are positive.

In our proof of Theorem 4.1 — the fact that, if R_Δ satisfies Bézout's Lemma, then Q represents either the norms or their negatives — we considered the set

$$J_Q = \{ax + \frac{b+\sqrt{\Delta}}{2}y : x,y\in\mathbb{Z}\}.$$

We showed that, if you multiply an element of J_Q by anything from R_Δ, the result is still in J_Q.

Subsets like J_Q are important enough to have a name: we'll say that a set $I\subseteq R_\Delta$ is an **ideal** if

- I is closed under addition, i.e., if $u,v\in I$ then $u+v\in I$.
- If $u\in I$ and $r\in R_\Delta$, then $ru\in I$.

Given some list of elements $u_1,\dots,u_n\in R_\Delta$, we can always form an ideal by considering all possible sums of different multiples of the u_i's. The common way to denote this ideal is to just wrap the list of elements in parentheses, that is,

$$(u_1,\dots,u_n) = \{r_1u_1+\cdots+r_nu_n : r_1,\dots,r_n\in R_\Delta\}.$$

The ideal J_Q that we got from our quadratic form can be written in this form quite simply: we have

$$J_Q = \left(a, \frac{b+\sqrt{\Delta}}{2}\right).$$

When an ideal I is generated by just one element, say $I = (d)$ for some $d \in R_\Delta$, we'll say that I is **principal**, and specifically call it the **principal ideal generated by** d. This is the same as just saying that I consists of all the multiples of d. I encourage you to convince yourself — possibly using some ideas from Exercise 2.3 — that if R_Δ satisfies Bézout's Lemma, then *every* ideal in R_Δ is principal. Rings with this property are called **principal ideal domains**.

Using this new language, we could say that what we showed in Theorem 4.1 was that, whenever J_Q is a principal ideal, Q represents either the norms or their negatives. If R_Δ is a principal ideal domain, this will always be true, but if not, then it will be true for some quadratic forms but not others.

Let's see how this shakes out in our examples in discriminant -20. The form x^2+5y^2 will give us the ideal $(1, \sqrt{-5})$, which is just the entirety of R_Δ. This ideal is certainly principal — it's generated by 1 — and sure enough, this form gives us the norms. I encourage you to compute that the form $2x^2+2xy+3y^2$ gives us the ideal $(2, 1+\sqrt{-5})$. Because this form doesn't represent the norms, this ideal can't be principal, which you could also verify directly if you like.

5.3 Ideal Products and Norms

Because we're interested in seeing what happens when we multiply the numbers represented by our respective forms, it'll be helpful to see what happens when we multiply elements from two *ideals* by each other. We'll do this by introducing a product operation on ideals. It might be tempting to define the product of two ideals I and J as something like $\{uv : u \in I, v \in J\}$, but this isn't quite right: this set is usually not going to be closed under addition, and therefore it won't be an ideal. To get an ideal, we instead look at all possible *sums* of elements of this form, that is,

$$IJ = \left\{ \sum_{i=1}^{n} u_i v_i : u_i \in I, v_i \in J \right\}.$$

For the ideal $J = (2, 1+\sqrt{-5})$, I encourage you to convince yourself that we get $J^2 = (4, 2+2\sqrt{-5}, -4+2\sqrt{-5}) = (2)$. In particular, notice that, even though J isn't principal, its *square* is! As we'll see momentarily, this is essentially the reason why the product of two elements represented by $2x^2+2xy+3y^2$ is always represented by x^2+5y^2.

Recall that the relationship between the ideal and the form was that

$$Q(x,y) = \frac{1}{a} N\left(ax + \frac{b+\sqrt{\Delta}}{2} y\right),$$

and that J_Q contains exactly the elements that show up inside the norm here. So knowing which ideal J_Q is almost tells you which numbers are represented by Q; the only obstacle is the factor of $1/a$ in front. If we could find a way to extract a just from the ideal, then we'd be in business.

Just as we defined norms of elements of R_Δ, it's in fact possible to define norms of ideals in a similar way: given an ideal $I \subseteq R_\Delta$, we can define $\overline{I} = \{\overline{u} : u \in I\}$. When I has the property that $I\overline{I} = (n)$ for some positive integer n, we'll say that I is **invertible** and define $N(I) = n$. In our computation earlier with $J = (2, 1+\sqrt{-5})$, it turns out that J is its own conjugate, and therefore $N(J) = 2$.

What if our ideal *isn't* invertible? It is also possible to define a norm in that case, but the theory is much less nice. It turns out that, if the coefficients a, b, c of the quadratic form Q have no common factors, then the ideal J_Q will always be invertible. In particular, you can check that, if Δ can't be written as a square times another discriminant, then *every* form of discriminant Δ will have coefficients with no common factors. Such Δ's are called **fundamental discriminants**, and for simplicity, we'll assume that Δ is fundamental from now on.

In fact, I encourage you to compute that, in this happy scenario, we have $N(J_Q) = a$.[10] This means we can in fact recover the numbers represented by our form just from the ideal J_Q: you can check that an integer n is represented by Q if and only if $n = N(u)/N(J_Q)$ for some $u \in J_Q$.

Suppose m and n are represented by $2x^2 + 2xy + 3y^2$. This means we have $m = N(u)/N(J)$ and $n = N(v)/N(J)$ for some $u, v \in J = (2, 1+\sqrt{-5})$. Therefore, $mn = N(uv)/N(J^2)$. (It's straightforward to check that, for any invertible ideals I, I', we have $N(II') = N(I)N(I')$.) But we showed earlier that $J^2 = (2)$, so $N(J^2) = 4$. Since $uv \in J^2 = (2)$, we get that $uv = 2w$ for some w, and therefore $mn = N(2w)/4 = N(w)$. In other words, the fact J^2 was principal does indeed tell us that mn is a norm.

5.4 The Ideal-Form Story in General

We're now ready to describe the relationship between quadratic forms and ideals in general, or at least whenever $\Delta < 0$, our form takes positive values, and Δ is a fundamental discriminant.

Suppose we start from a form Q. We've already seen how to produce an ideal J_Q, and our procedure also gave us a pair of elements $\alpha, \beta \in J_Q$ for which $J_Q = (\alpha, \beta)$ and

$$Q(x, y) = \frac{N(x\alpha + y\beta)}{N(J_Q)}.$$

(Specifically, $\alpha = a$ and $\beta = \frac{b+\sqrt{\Delta}}{2}$.)

This also gives us a nice way to go backwards, from ideals to forms. If I is an ideal in R_Δ, it is in fact always possible to write $I = \{\alpha x + \beta y : x, y \in \mathbb{Z}\}$. We won't prove this, but if you pick such a pair of generators which is *positively oriented* — that is, the angle from α to β in the complex plane is counter-clockwise — we can produce the quadratic form

$$Q_I(x, y) = \frac{N(x\alpha + y\beta)}{N(I)}.$$

What happens if we replace Q by an equivalent form and look at J_Q? You might expect from the way we've been setting things up that the ideal J_Q will be unchanged, but this is not quite true. For example, the form $2x^2 + 2xy + y^2$ of discriminant -4 is equivalent to $x^2 + y^2$, but it gives the ideal $(2, 1+i) = (1+i)$, whereas $x^2 + y^2$ gives the ideal (1).

What *is* true is that, if Q and Q' are equivalent via a change of coordinates with determinant 1 (rather than -1), then it will always be the case that $r J_Q = r' J_{Q'}$ for some $r, r' \in R_\Delta$. When the change of coordinates between Q and Q' has determinant 1, we'll say Q and Q' are **properly equivalent**. The procedure for going from ideals to forms we just described doesn't in fact care if you multiply the ideal by a constant: we always have $N(rI) = N(r)N(I)$, and I encourage you to check that this means Q_I and Q_{rI} are the same quadratic form (if you choose the right α and β both times).

[10] If we didn't assume a was positive, this would have to say $|a|$.

Whenever two ideals I and I' have the property that $rI = r'I'$ for some $r, r' \in R_\Delta$, we'll say that I and I' are **equivalent**. The big theorem is then as follows: *If $\Delta < 0$ is a fundamental discriminant, there is a one-to-one correspondence between quadratic forms of discriminant Δ with positive values up to proper equivalence and nonzero ideals of R_Δ up to equivalence. Furthermore, if m is represented by a form corresponding to an ideal I and n is represented by a form corresponding to an ideal J, then mn is represented by a form corresponding to the ideal IJ.*

(What happens if instead Q and Q' are related by a change of coordinates with determinant -1? In this case, it turns out that J_Q is equivalent to $\overline{J_{Q'}}$. If J_Q isn't equivalent to its own conjugate, that means we'll get a different ideal out, which will get in the way of building a nice one-to-one correspondence. For this reason, we'll restrict our attention to proper equivalence of forms from now on.)

The set of equivalence classes of nonzero ideals under multiplication forms what's called an **abelian group**. This means multiplication is associative, it has an identity element (namely the class containing the ideal (1)), and every ideal has an inverse — the inverse of I is in fact just $\overline{I}$, since $I\overline{I}$ is principal and therefore equivalent to (1). This group is called the **ideal class group** of R_Δ, and it's one of the most important structures in algebraic number theory.

How could we turn this into a procedure for determining which numbers are represented by which forms of a given discriminant? For this we'll need one more big fact. If Δ is a fundamental discriminant, then there's a unique factorization theorem for ideals in R_Δ. That is, ideals in R_Δ can be factored into "irreducibles," called **prime ideals**, and these factorizations are unique. This is true whether or not R_Δ has unique factorization of *elements*![11]

From this fact, it's possible to prove a classification of prime ideals in R_Δ that's very similar to our classification of irreducibles in $\mathbb{Z}[i]$: a prime integer p is **inert** if (p) is a prime ideal in R_Δ; the only other possibility is that $(p) = P\overline{P}$ for some prime ideal P, and in this case we say p is **ramified** if $P = \overline{P}$ and **split** if $P \neq \overline{P}$. Like in the Gaussian integers, every prime ideal appears in this way: if p is inert, there's one prime ideal of norm p^2; if p is ramified, there's one prime ideal of norm p; and if p is split, then there are two prime ideals of norm p which are conjugates of each other.

The procedure for determining whether n is represented by some form Q of discriminant Δ then shakes out as follows. In order to be represented in discriminant Δ, each prime that isn't represented in discriminant Δ needs to appear an even number of times. Each prime p that *is* represented in discriminant Δ will be either ramified or split, that is, $(p) = P\overline{P}$ for some prime ideal P. The only forms which represent p will turn out to be the ones corresponding to the ideals P and $\overline{P}$, and from here we can produce a list of all the forms representing n: if $n = d^2 p_1^{a_1} \cdots p_n^{a_n}$, where each $p_i = P_i\overline{P_i}$, then a form Q represents n if and only if the class of the corresponding ideal in the class group can be written as $P_1^{\pm a_1} \cdots P_n^{\pm a_n}$, where the minus sign in an exponent means we take the conjugate of the corresponding ideal.

[11] In fact, this is the original reason for the name "ideal": they're kind of like numbers in R_Δ, but they're better, because they have unique factorization even if the numbers don't.