representations of 5 and 13 as sums of two squares by Diophantus' identity

$$(a_1^2+b_1^2)(a_2^2+b_2^2)=(a_1a_2\pm b_1b_2)^2+(b_1a_2\mp a_1b_2)^2.$$

Fermat claimed to have proved that each prime of the form $4n+1$ is a sum of two squares in exactly one way. However, his proof was lost, and the first known proof was given by Euler in 1749. Euler's proof was heavy going, and the theorem became a challenge to later mathematicians, to test the strength of new methods in number theory. Progressively easier and more elegant proofs were given by Lagrange, Gauss, and Dedekind. The following proof uses ideas from all three, but the crux of it is uniqueness of Gaussian prime factorization.

Fermat's two squares theorem. *If p is a prime of the form $4n+1$, then $p=a^2+b^2$ for a unique pair of natural numbers a and b.*

Proof The first step is to find a square, m^2, such that p divides m^2+1. Lagrange found a way to do this using Wilson's theorem.[2] Recall from Section 6.5 that this theorem says $(p-1)!\equiv -1 \pmod p$ when p is prime, so when $p=4n+1$ we have

$$\begin{aligned}-1 &\equiv 1\times 2\times\cdots\times 4n \pmod p\\ &\equiv (1\times 2\times\cdots\times 2n)(2n+1)\times\cdots\times(4n-1)\times 4n \pmod p\\ &\equiv (1\times 2\times\cdots\times 2n)(-2n)\times\cdots\times(-2)\times(-1) \pmod p\\ &\quad\text{because each } p-k\equiv -k \pmod p\\ &\equiv (1\times 2\times\cdots\times 2n)^2(-1)^{2n} \pmod p\\ &\equiv (1\times 2\times\cdots\times 2n)^2 \pmod p.\end{aligned}$$

Thus if we take $m=(2n)!$ we have $-1\equiv m^2 \pmod p$, and therefore p divides m^2+1.

Now m^2+1 has the Gaussian integer factorization $(m+i)(m-i)$, and p does *not* divide $m+i$ or $m-i$, because the quotients $\frac{m}{p}+\frac{i}{p}$ and $\frac{m}{p}-\frac{i}{p}$ are not Gaussian integers. It then follows from unique Gaussian prime factorization (or, more particularly, from the Gaussian prime

[2] The idea was indicated in Exercise 6.5.9, but to avoid dependence on the exercises, the details are given here.

divisor property) that *p is not a Gaussian prime.* Thus p has a Gaussian prime divisor, say, $a+ib$. Now

$$\begin{aligned} a+ib \text{ divides } p \Rightarrow p &= (a+ib)c \\ &\quad \text{for some nonunit Gaussian integer } c \\ \Rightarrow p &= (a-ib)\bar{c} \quad \text{taking conjugates of both sides} \\ \Rightarrow p^2 &= (a^2+b^2)|c|^2 \\ &\quad \text{multiplying preceding equations} \\ \Rightarrow p &= a^2+b^2 \\ &\quad \text{by unique prime factorization} \\ &\quad \text{of natural numbers.} \end{aligned}$$

Conversely, if a prime $p = a^2+b^2$, then p has the Gaussian factorization $p = (a+ib)(a-ib)$, and each factor is a Gaussian prime, because its norm is the prime p. Thus a and b are the real and imaginary parts (up to sign) of the unique Gaussian prime factors of p, and therefore a^2+b^2 is the unique sum of squares equal to p. □

This theorem tells us all the primes that are sums of two squares. Apart from those of the form $4n+1$, the only such prime is $2 = 1^2+1^2$, because primes of the form $4n+3$ are not sums of two squares, by a congruence mod 4 argument.

The proof also tells us that if an ordinary prime p is not a Gaussian prime, then $p = a^2+b^2$. Hence *ordinary primes of the form* $4n+3$ *are Gaussian primes*. This information leads to the following theorem, which is closely allied with Fermat's two squares theorem and unique Gaussian prime factorization. It shows that the Gaussian primes can be regarded as "known" once the ordinary primes are known.

Classification of Gaussian primes. *Up to unit factors, the Gaussian primes are*

- *Ordinary primes of the form* $4n+3$.
- *The factors* $a+ib$, $a-ib$ *of primes* a^2+b^2 *of the form* $4n+1$ *or* 2.

Proof By the preceding remarks, the only ordinary primes that are Gaussian primes are those of the form $4n+3$. The factors $a+ib$ and $a-ib$ of primes $p = a^2+b^2$ are Gaussian primes because they have norm p.

Conversely, if $a+ib$ is a Gaussian prime with $a, b \neq 0$ then so is its conjugate $a-ib$, because a nontrivial factorization of $a-ib$ would give one of $a+ib$ by conjugation. Thus Gaussian primes that are not ordinary primes come in pairs $a \pm ib$. The product a^2+b^2 of such a pair is an ordinary prime, by unique Gaussian prime factorization, because a factorization of a^2+b^2 into ordinary primes would be different from its Gaussian prime factorization $(a+ib)(a-ib)$. Such primes are 2 and those of the form $4n+1$ by Fermat's two squares theorem. □

Exercises

It is possible to study the primes of $\mathbb{Z}[\sqrt{-2}]$ in a similar way, using its unique prime factorization theorem from the previous exercise set. However, to do something a little different, we shall use $\mathbb{Z}[\sqrt{-2}]$ to investigate the equation $y^3 = x^2+2$. Diophantus mentioned the natural number solution $x=5$, $y=3$ to this equation, and Fermat claimed it was the only one. The first known proof was given by Euler (1770), assuming unique prime factorization in $\mathbb{Z}[\sqrt{-2}]$ (but failing to mention it). Such a proof can be carried out rigorously as follows.

Assuming x are natural numbers with $y^3 = x^2+2$, note that

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

This transforms the problem into one about cubes in $\mathbb{Z}[\sqrt{-2}]$.

7.6.1. Use congruences mod 4 to show that x is odd.

7.6.2. Deduce from Exercise 7.6.1 that $\gcd(x+\sqrt{-2}, x-\sqrt{-2}) = 1$.

Thus we have relatively prime numbers $x+\sqrt{-2}$ and $x-\sqrt{-2}$ whose product is a cube, y^3. We can conclude that $x+\sqrt{-2}$ and $x-\sqrt{-2}$ are themselves cubes in $\mathbb{Z}[\sqrt{-2}]$ by the remarks at the end of Section 7.5 and the fact that units of $\mathbb{Z}[\sqrt{-2}]$ are ± 1 (Exercise 7.4.5).

7.6.3. Suppose $x+\sqrt{-2} = (a+b\sqrt{-2})^3$ is a cube in $\mathbb{Z}[\sqrt{-2}]$, so a and b are ordinary integers. Deduce that

$$x = a^3 - 6ab^2 = a(a^2-6b^2) \quad \text{and} \quad 1 = 3a^2b - 2b^3 = b(3a^2-2b^2).$$

7.6.4. Deduce from Exercise 7.6.3 that $b = \pm 1$, $a = \pm 1$ and hence the only natural number solution for x is 5.

7.7* Factorizing a Sum of Two Squares

Diophantus' identity

$$(a^2 + b^2)(c^2 + d^2) = (ac - bd)^2 + (ad + bc)^2$$

tells us that when we multiply sums of two squares the product is also a sum of two squares. What happens when we divide? Is a divisor of a sum of two squares a sum of two squares? If a and b have a common divisor d, then $a^2 + b^2$ has the divisor d^2, which is trivially the sum $0^2 + d^2$ of two integer squares. And if we stick to a and b with no common prime divisor we have the following elegant theorem, discovered by Euler in 1747.

Divisors of sums of two squares. *If* $\gcd(a, b) = 1$, *then any divisor of* $a^2 + b^2$ *is of the form* $c^2 + d^2$, *where* $\gcd(c, d) = 1$.

Proof Each divisor $e > 1$ of $a^2 + b^2$ is a product of *Gaussian* prime divisors of $a^2 + b^2$, by unique prime factorization in $\mathbb{Z}[i]$. Because $a^2 + b^2 = (a + ib)(a - ib)$, each Gaussian prime divisor $q + ir$ of $a^2 + b^2$ divides either $a + ib$ or $a - ib$. And because $\gcd(a, b) = 1$, none of the divisors $q + ir$ is a real prime p, as $\frac{a}{p} \pm i\frac{b}{p}$ is not in $\mathbb{Z}[i]$.

Now the Gaussian prime divisors of e occur in conjugate pairs $q + ir$, $q - ir$, because if $q + ir$ divides e so does $q - ir$, by taking conjugates. From each pair we collect the member dividing $a + ib$, and form their product $c + id$. Then the conjugate members dividing $a - ib$ have product $c - id$, and

$$e = (c + id)(c - id) = c^2 + d^2.$$

Also $\gcd(c, d) = 1$, because a common (real) prime divisor p of c and d would divide $a + ib$ and hence both a and b, contrary to assumption. □

This proof is another example of the way $\mathbb{Z}[i]$ refines our understanding of $\mathbb{Z}$. It shows that factorization into natural numbers of the form $x^2 + y^2$ can be viewed as a consequence of factorization into Gaussian integers. Moreover, it is *simpler* to view the situation this way, as the proof using real integers alone is more complicated.

Exercises

There is a similar theorem about divisors of numbers of the form $a^2 + 2b^2$, and it may be proved similarly using unique prime factorization in $\mathbb{Z}[\sqrt{-2}]$.

7.7.1. If $\gcd(a, b) = 1$, show that any divisor of $a + b\sqrt{-2}$ is of the form $c + d\sqrt{-2}$ with $\gcd(c, d) = 1$.

7.7.2. Deduce from Exercise 7.7.1 that if $\gcd(a, b) = 1$ then any divisor of $a^2 + 2b^2$ is of the form $c^2 + 2d^2$ with $\gcd(c, d) = 1$.

Euler's theorem on the divisors of a^2+b^2 ties up nicely with the idea of "factorizing" Pythagorean triples explored in the exercises to Section 5.4.

7.7.3. Show that a divisor $c^2 + d^2$ of $a^2 + b^2$ corresponds to a Pythagorean triple $(2cd, b^2 - c^2, b^2 + c^2)$, which is a "factor" of the triple $(2ab, a^2 - b^2, a^2 + b^2)$. Illustrate this result with the triple (319, 360, 481) from Plimpton 322.

7.8 Discussion

Complex Numbers and Geometry

The geometry of complex numbers is a vast subject. It covers not only the Euclidean plane but also the sphere, the non-Euclidean plane and even non-Euclidean space. On all of these objects, it is possible to describe isometries by simple functions of a complex variable. We have seen how this happens when the Euclidean plane is interpreted as $\mathbb{C}$, and it happens similarly on the sphere and the non-Euclidean plane when they are suitably mapped to the Euclidean plane.

A complex coordinate on the sphere is obtained by stereographic projection (Section 4.6*) from the sphere to the plane $\mathbb{C}$. This projects every point on the sphere, except the north pole, to a complex number z we take as its coordinate. We take ∞ as the coordinate of the north pole, which works perfectly in this situation. In particular, the half turn of the sphere about the real axis that exchanges the north and south poles sends the point with coordinate z to the point with coordinate $1/z$, so ∞ is exchanged with $1/\infty = 0$, as it should be. General rotations of the sphere turn out to be the functions of the form $\frac{az+b}{-\bar{b}z+\bar{a}}$.

The non-Euclidean plane mentioned in Section 3.9* has an obvious complex coordinate, because it is naturally viewed as the half plane of complex numbers $x + iy$ with $y > 0$. In fact, this is how it was introduced by Poincaré (1882), and he went on to show that its orientation-preserving isometries are the functions $\frac{az+b}{cz+d}$ with a, b, c, d real and $ad - bc > 0$. Examples are the function $2z$, which "translates" points z along the imaginary axis, and $-1/z$, which is a "half turn" about the point i. The simplest orientation-reversing isometry is reflection $-\bar{z}$ of z in the imaginary axis.

Figure 7.6 shows a tessellation of the half plane by triangles with angles $\pi/2$, $\pi/3$, and $\pi/7$. It is clear enough to the eye that all the triangles have the same angles, but they are also *congruent* in the sense of non-Euclidean geometry. The picture was in fact generated from one triangle by repeatedly reflecting in its sides. One of its symmetries is a translation along the imaginary axis.

Projection from line to circle (Section 4.6*) can be extended to a map $f(z) = \frac{z-i}{z+i}$ of the half plane onto the unit disk $\{z : |z| < 1\}$. (Incidentally, this accounts for the formula $\frac{t-i}{t+i}$ in Exercise 4.3.2 that gives all rational points on the circle as t runs through the rationals.) This correspondence between the half plane and the disk allows the latter to be used as another "model" of the non-Euclidean plane, much as one uses different map projections in geography to model the sphere. The half plane and disk models look somewhat similar, because the function $f(z) = \frac{z-i}{z+i}$ preserves angles and circles.

The disk model also reveals a striking algebraic analogy between the sphere and the non-Euclidean plane. When we use z as a coordinate in the disk model, its orientation-preserving isometries are

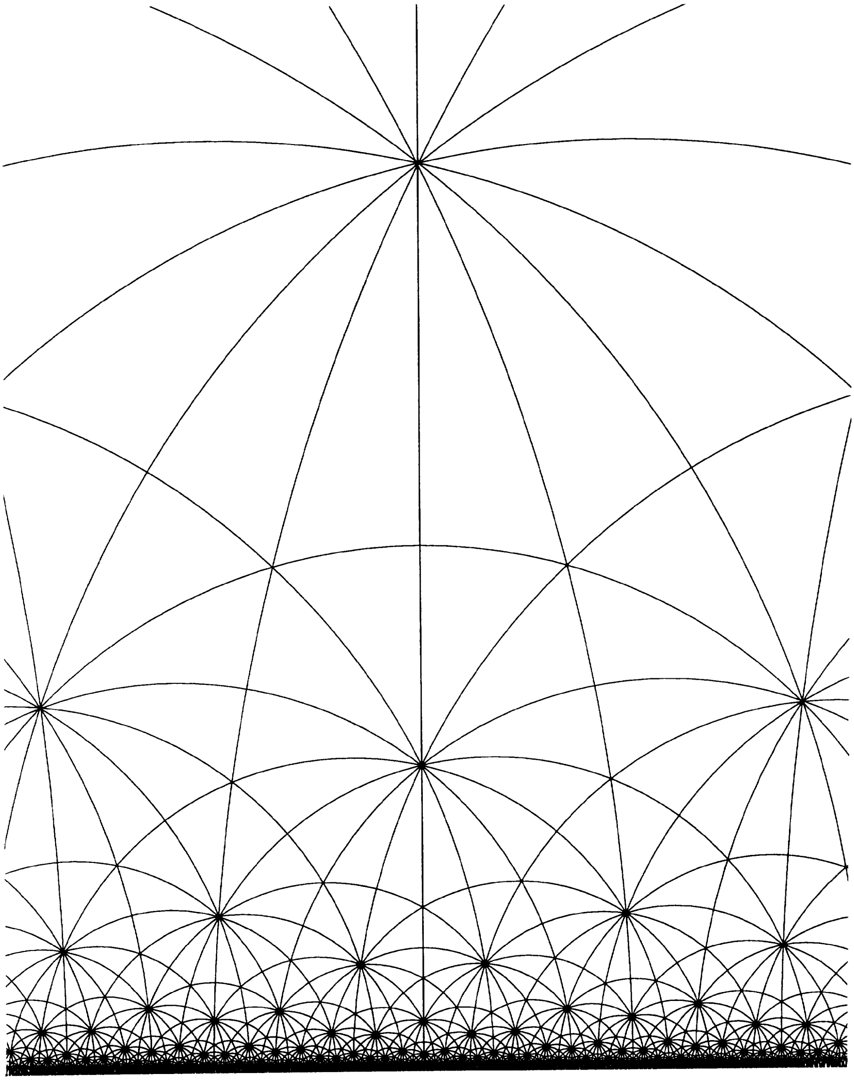

FIGURE 7.6 (2,3,7) tessellation of the half plane.