Part IV
Branches of Mathematics

IV.1 Algebraic Numbers

Barry Mazur

T&T note: check style later.

The roots of our subject go back to ancient Greece while its branches touch almost all aspects of contemporary mathematics. In 1801 the *Disquisitiones Arithmeticae* of CARL FRIEDRICH GAUSS [VI.26] was first published, a "founding treatise," if ever there was one, for the modern attitude toward number theory. Many of the still unachieved aims of current research can be seen, at least in embryonic form, as arising from Gauss's work.

This article is meant to serve as a companion to the reader who might be interested in learning, and thinking about, some of the classical theory of algebraic numbers. Much can be understood, and much of the beauty of algebraic numbers can be appreciated, with a minimum of theoretical background. I recommend that readers who wish to begin this journey carry in their backpacks Gauss's *Disquisitiones Arithmeticae* as well as Davenport's *The Higher Arithmetic* (1992), which is one of the gems of exposition of the subject, and which explains the founding ideas clearly and in depth using hardly anything more than high-school mathematics.

1 The Square Root of 2

The study of algebraic numbers and algebraic integers begins with, and constantly reverts back to, the study of ordinary rational numbers and ordinary integers. The first algebraic irrationalities occurred not so much as *numbers* but rather as *obstructions* to simple answers to questions in geometry.

That the ratio of the diagonal of a square to the length of its side cannot be expressed as a ratio of whole numbers is purported to be one of the vexing discoveries of the early Pythagoreans. But this very ratio, when squared, is 2:1. So we might—and later mathematicians certainly did—deal with it algebraically. We can think of this ratio as a cipher, about which we know nothing beyond the fact that its square is 2 (a viewpoint taken toward algebraic numbers by KRONECKER [VI.48], as we shall see below). We can write $\sqrt{2}$ in various forms, e.g.,

$$\sqrt{2} = |1 - \mathrm{i}|, \tag{1}$$

and we can think of $1 - \mathrm{i} = 1 - e^{2\pi \mathrm{i}/4}$ as the world's simplest trigonometric sum; we shall see generalizations of this for all quadratic surds below. We can also view $\sqrt{2}$ as a limit of various infinite sequences, one of which is given by the elegant CONTINUED FRACTION [III.22]

$$\sqrt{2} = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \ddots}}. \tag{2}$$

Directly connected to this continued fraction (2) is the Diophantine equation

$$2X^2 - Y^2 = \pm 1 \tag{3}$$

known as the *Pell equation*. There are infinitely many pairs of integers (x, y) satisfying this equation, and the corresponding fractions y/x are precisely what you get by truncating the expression in (2). For example, the first few solutions are $(1, 1)$, $(2, 3)$, $(5, 7)$, and $(12, 17)$, and

$$\left.\begin{aligned} \tfrac{3}{2} &= 1 + \tfrac{1}{2} = 1.5, \\ \tfrac{7}{5} &= 1 + \cfrac{1}{2 + \tfrac{1}{2}} = 1.4, \\ \tfrac{17}{12} &= 1 + \cfrac{1}{2 + \cfrac{1}{2 + \tfrac{1}{2}}} = 1.416\ldots. \end{aligned}\right\} \tag{4}$$

Replace the ± 1 on the right-hand side of (3) by *zero* and you get $2X^2 - Y^2 = 0$, an equation all of whose positive real-number solutions (X, Y) have the ratio $Y/X = \sqrt{2}$, so it is easy to see that the sequence of fractions (4) (these being alternately larger and smaller than $\sqrt{2} = 1.414\ldots$) converges to $\sqrt{2}$ in the limit. Even more striking is that (4) is a list of fractions that best approximate $\sqrt{2}$. (A rational number a/d is said to be a *best approximant* to a real number α if a/d is closer to α than any rational number of denominator smaller than or equal to d.) To deepen the pic-

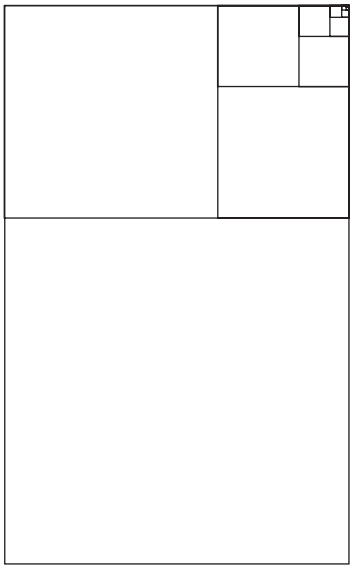

Figure 1 The outer rectangle has its height-to-width ratio equal to the golden mean. If you remove a square from it as indicated in the figure, you are left with a rectangle that has the golden mean as its width-to-height ratio. This procedure is of course repeatable.

ture, consider another important infinite expression, the conditionally convergent series

$$\frac{\log(\sqrt{2}+1)}{\sqrt{2}} = 1 - \tfrac{1}{3} - \tfrac{1}{5} + \tfrac{1}{7} + \tfrac{1}{9} + \cdots \pm \tfrac{1}{n} + \cdots . \quad (5)$$

Here the n range over positive odd numbers, and the sign of the term $\pm 1/n$ is *plus* if n has a remainder of 1 or 7 when divided by 8, and it is *minus* if n has a remainder of 3 or 5. This elegant formula (5), which you are invited to "check out" at least to one digit accuracy with a calculator, is an instance of the powerful and general theory of *analytic formulas for special values of L*-FUNCTIONS [III.49], which plays the role of a bridge between the more algebraic and the more analytic sides of the story. When we allude to this, below, we will call it "the analytic formula," for short.

PUP: a comma before this 'is' would not be good.

2 The Golden Mean

If you are looking for quadratic irrationalities that have been the subject of geometric fascination through the ages, then $\sqrt{2}$ has a strong competitor in the number $\frac{1}{2}(1+\sqrt{5})$, known as the *golden mean*. The ratio $\frac{1}{2}(1+\sqrt{5})$:1 gives the proportions of a rectangle with the property that when you remove a square from it, as in figure 1, you are left with a smaller rectangle whose sides are in the same proportion. Its corresponding trigonometric sum description is

$$\tfrac{1}{2}(1+\sqrt{5}) = \tfrac{1}{2} + \cos\tfrac{2}{5}\pi - \cos\tfrac{4}{5}\pi. \quad (6)$$

Its continued-fraction expansion is

$$\tfrac{1}{2}(1+\sqrt{5}) = 1 + \cfrac{1}{1+\cfrac{1}{1+\ddots}}, \quad (7)$$

where the sequence of fractions obtained by successive truncations of this continued fraction,

$$\frac{y}{x} = \tfrac{1}{1}, \tfrac{2}{1}, \tfrac{3}{2}, \tfrac{5}{3}, \tfrac{8}{5}, \tfrac{13}{8}, \tfrac{21}{13}, \tfrac{34}{21}, \ldots, \quad (8)$$

is a sequence of best rational-number approximants to

$$\tfrac{1}{2}(1+\sqrt{5}) = 1.618033988749894848\ldots,$$

where "best" has the sense already mentioned. For example, the fraction

$$\frac{34}{21} = 1 + \cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\cfrac{1}{1+\frac{1}{1}}}}}}$$

equals $1.619047619047619047\ldots$ and is closer to the golden mean than any fraction with denominator less than 21.

Nevertheless, the exclusive appearance of 1s in this continued fraction[1] can be used to show that, among all irrational real numbers, the golden mean is the number that is, in a specific technical sense, *least well approximated by rational numbers*.

Readers familiar with the sequence of Fibonacci numbers will recognize them in the successive denominators of (8), and in the numerators as well. The analogue to equation (3) is

$$X^2 + XY - Y^2 = \pm 1. \quad (9)$$

PUP: big thanks to proofreader for spotting incorrect equation cross-references in this article.

This time, if you replace the ± 1 on the right-hand side of the equation by 0, you get the equation $X^2 + XY - Y^2 = 0$, whose positive real-number solutions (X, Y) have the ratio $Y/X = \frac{1}{2}(1+\sqrt{5})$, that is, the golden mean. And now the numerators and denominators y, x that appear in (8) run through the positive integral solutions of (9). The analogue of formula (5) (the "analytic formula") for the golden mean is the conditionally convergent infinite sum

$$\frac{2\log(\frac{1}{2}(1+\sqrt{5}))}{\sqrt{5}} = 1 - \tfrac{1}{2} - \tfrac{1}{3} + \tfrac{1}{4} + \tfrac{1}{6} + \cdots \pm \tfrac{1}{n} + \cdots , \quad (10)$$

where the n range over positive integers not divisible by 5, and the sign of $\pm 1/n$ is *plus* if n has a remainder

PUP: again, the proofreader's suggested comma would definitely not be a good addition before this 'is'.

1. The continued-fraction expansion of any real quadratic algebraic number has an eventually recurring pattern in its entries, as is vividly exhibited by the two examples (2) and (7) given above.

of ± 1 when divided by 5, and *minus* otherwise.

What governs the choice of the plus terms and minus terms is whether or not n is a *quadratic residue modulo* 5. Here is a brief explanation of this terminology. If m is an integer, two integers a, b are said to be *congruent modulo* m (in symbols we write $a \equiv b \mod m$) if the difference $a - b$ is an integral multiple of m; if a, b, and m are positive numbers, it is equivalent to ask that a and b have the same "remainder" (sometimes also called "residue") when each is divided by m (see MODULAR ARITHMETIC [III.60]). An integer a relatively prime to m is called a *quadratic residue* modulo m if a is congruent to the square of some integer, modulo m; otherwise it is called a *quadratic nonresidue* modulo m. So, $1, 4, 6, 9, \ldots$ are quadratic residues modulo 5, while $2, 3, 7, 8, \ldots$ are quadratic nonresidues modulo 5.

A generalization of equations (5) and (10) (the "analytic formula for the L-function attached to quadratic Dirichlet characters") gives a very surprising formula for the conditionally convergent sum of terms $\pm 1/n$, where n runs through positive integers relatively prime to a fixed integer and the sign of $\pm 1/n$ corresponds to whether n is a quadratic residue, or nonresidue modulo that integer.

3 Quadratic Irrationalities

The quadratic formula

$$X = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

gives the solutions (usually two) to the general quadratic polynomial equation $aX^2 + bX + c = 0$ as a rational expression of the number $\sqrt{D}$, where $D = b^2 - 4ac$ is known as the *discriminant* of the polynomial $aX^2 + bX + c$, or, equivalently, of the corresponding homogeneous QUADRATIC FORM [III.75] $aX^2 + bXY + cY^2$. This formula introduces many irrational numbers: Plato's dialogue "Theaetetus" has the young Theaetetus credited with the discovery that $\sqrt{D}$ is irrational whenever D is a natural number that is not a perfect square. The curious switch, from initially perceiving an *obstruction* to a problem to eventually embodying this obstruction as a *number* or an *algebraic object of some sort* that we can effectively study, is repeated over and over again, in different contexts, throughout mathematics. Much later, *complex* quadratic irrationalities also made their appearance. Again these were not at first regarded as "numbers as such," but rather as *obstructions* to the solution of problems. Nicholas Chuquet, for example, in his 1484 manuscript, *Le Triparty*, raised the question of whether or not there is a number whose triple is four plus its square and he comes to the conclusion that there is no such number because the quadratic formula applied to this problem yields "impossible" numbers, i.e., complex quadratic irrationalities in our terminology.[2]

For any real quadratic ("integral") irrationality there is a discussion along similar lines to the ones we have just given (expressions (1)–(5) for $\sqrt{2}$ and expressions (6)–(10) for $\frac{1}{2}(1 + \sqrt{5})$). For complex irrationalities, there is also such a theory, but with interesting twists. For one thing, we do not have anything directly comparable to continued-fraction expansions for a complex quadratic irrationality. In fact, the simple, but true, answer to the problem of how to find an infinite number of rational numbers that converge to such an irrationality is that you cannot! Correspondingly, the analogue of the Pell equation has only finitely many solutions. As a consolation, however, the appropriate "analytic formula" has a simpler sum, as we will see below.

Let d be any square-free integer, positive or negative. Associated with d is a particularly important number τ_d, defined as follows. If d is congruent to 1 mod 4 (that is, if $d - 1$ is a multiple of 4), then $\tau_d = \frac{1}{2}(1 + \sqrt{d})$; otherwise, $\tau_d = \sqrt{d}$. We will refer to these quadratic irrationalities τ_d as *fundamental algebraic integers of degree* 2. The general notion of an "algebraic integer" is defined in section 11. An algebraic integer of degree two is simply a root of a quadratic polynomial of the form $X^2 + aX + b$ with a, b ordinary integers. In the first case (when $d \equiv 1$ modulo 4), τ_d is a root of the polynomial $X^2 - X + \frac{1}{4}(1 - d)$ and in the second it is a root of $X^2 - d$. The reason special names are given to these quadratic irrationalities is that *any* quadratic algebraic integer is a linear combination (with ordinary integers as coefficients) of 1 and one of these fundamental quadratic algebraic integers.

4 Rings and Fields

I think that one of the big early advances in mathematics is the now-current, universal recognition of the importance of studying the properties of *collections* of mathematical objects, and not just the objects in isolation. A *ring* R of complex numbers is a collection of

2. BOMBELLI [VI.8], in the sixteenth century, would refer to irrational square roots, of positive or of negative numbers, as "deaf" (reminiscent of the word *surd* that is still in use) and as "numbers impossible to name."

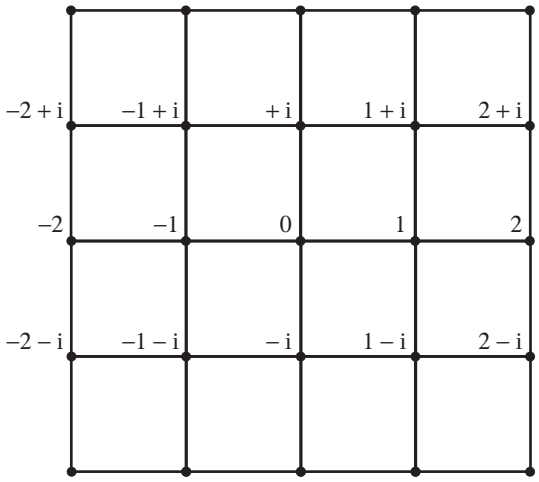

Figure 2 The Gaussian integers are the vertices of this lattice of squares tiling the complex plane.

them that contains 1 and is closed under the operations of addition, subtraction, and multiplication. That is, if a, b are any two numbers in R, $a \pm b$ and ab must also be in R. If such a ring R has the further property that it is closed under division by nonzero elements (i.e., if a/b is again in R whenever a and b are, and $b \neq 0$), then we say that R is a *field*. (These concepts are discussed further in FIELDS [I.3 §2.2] and RINGS, IDEALS, AND MODULES [III.83].) The ring $\mathbb{Z}$ of ordinary integers, $\{0, \pm 1, \pm 2, \dots\}$ is our "founding example" of a ring; visibly, it is the smallest ring of complex numbers.

The collection of all real or complex numbers that are integral linear combinations of 1 and τ_d is closed under addition, subtraction, and multiplication, and is therefore a ring, which we denote by R_d. That is, R_d is the set of all numbers of the form $a + b\tau_d$ where a and b are ordinary integers. These rings R_d are our first, basic, examples of *rings of algebraic integers* beyond that prototype, $\mathbb{Z}$, and they are the most important rings that are receptacles for quadratic irrationalities. Every quadratic irrational algebraic integer is contained in exactly one R_d.

For example, when $d = -1$ the corresponding ring R_{-1}, usually referred to as the ring of *Gaussian integers*, consists of the set of complex numbers whose real and imaginary parts are ordinary integers. These complex numbers may be visualized as the vertices of the infinite tiling of the complex plane by squares whose sides have length 1 (see figure 2).

When $d = -3$ the complex numbers in the corresponding ring R_{-3} may be visualized as the vertices of

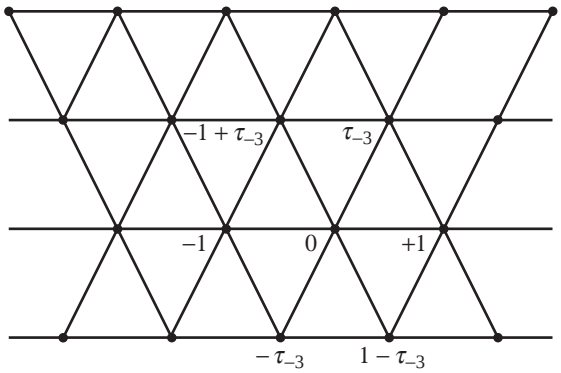

Figure 3 The elements of the ring R_{-3} are the vertices of this lattice of hexagons tiling the complex plane.

the regular hexagonal tiling of the complex plane (see figure 3).

With the rings R_d in hand, we may ask ring-theoretic questions about them, and here is some of the standard vocabulary useful for this. A *unit* u in a given ring R of complex numbers is a number in R whose reciprocal $1/u$ is also in R; a *prime* (or synonymously, an *irreducible*) element in R is a nonunit that cannot be written as the product of two nonunits in R. A ring of complex numbers R has the *unique factorization property* if every nonzero, nonunit, algebraic number in R can be expressed as a product of prime elements in exactly one way (where two factorizations are counted as the same if one can be obtained from the other by rearranging the order in which the primes appear and multiplying them by units).

In the prototype ring $\mathbb{Z}$ of ordinary integers, the only units are ± 1. The fundamental fact that any ordinary integer greater than 1 can be uniquely expressed as a product of (positive) prime numbers (that is, that $\mathbb{Z}$ enjoys the unique factorization property) is crucial for much of the number theory done with ordinary integers. That this unique factorization property for integers actually required proof was itself a hard-won realization of Gauss, who also provided its proof (see THE FUNDAMENTAL THEOREM OF ARITHMETIC [V.16]).

It is easy to see that there are only four units in the ring R_{-1} of Gaussian integers, namely ± 1 and $\pm i$; multiplication by any of these units effects a *symmetry* of the infinite square tiling (figure 2 above). There are only six units in the ring R_{-3}, namely ± 1, $\pm\frac{1}{2}(1 + \sqrt{-3})$ and $\pm\frac{1}{2}(1 - \sqrt{-3})$; multiplication by any of these units results in a symmetry of the infinite hexagonal tiling (figure 3 above).

Fundamental to understanding the arithmetic of R_d is the following question: which ordinary prime numbers p remain prime in R_d and which ones factorize into products of primes in R_d? We will see shortly that if a prime number does factorize in R_d, it must be expressible as the product of precisely two prime factors. For example, in the ring of Gaussian integers, R_{-1}, we have the factorizations

$$\begin{aligned} 2 &= (1+\mathrm{i})(1-\mathrm{i}),\\ 5 &= (1+2\mathrm{i})(1-2\mathrm{i}),\\ 13 &= (2+3\mathrm{i})(2-3\mathrm{i}),\\ 17 &= (1+4\mathrm{i})(1-4\mathrm{i}),\\ 29 &= (2+5\mathrm{i})(2-5\mathrm{i}),\\ &\vdots \end{aligned}$$

PUP: again we'd like to keep 'brackets' here, instead of changing to 'parentheses'. OK?

where all the Gaussian integer factors in brackets above are *prime* in the ring of Gaussian integers.

Let us say that an odd prime p *splits* in R_{-1} if it factorizes into a product of at least two primes and *remains prime* if it does not do so. As we shall soon see, the officially agreed-upon definitions of splitting and remaining prime for more general rings of algebraic integers (even ones of the form R_d) are worded slightly, but very significantly, differently from the way we have just defined these concepts in the ring R_{-1} of Gaussian integers. (Note that we have excluded the prime $p = 2$ from the above dichotomy. This is because 2 *ramifies* in R_{-1}; for a discussion of this concept see section 7 below.) In any event, there is an elementary computable *rule* that tells us, for any R_d, which primes p split and which remain prime in this agreed sense. The rule depends upon the residue of p modulo $4d$: the reader is invited to guess it for the ring of Gaussian integers given the data just displayed above. In general, an elementary computable rule that says which primes split and which do not in a ring of algebraic integers such as R_d is referred to as a *splitting law* for the ring of algebraic integers in question.

5 The Rings R_d of Quadratic Integers

There is a very important "symmetry," or AUTOMORPHISM [I.3 §4.1], defined on the ring R_d. It sends $\sqrt{d}$ to $-\sqrt{d}$, keeps all ordinary integers fixed, and more generally, for rational numbers u and v, it sends $\alpha = u + v\sqrt{d}$ to what we may call its *algebraic conjugate* $\alpha' = u - v\sqrt{d}$. (The word "algebraic" is to remind you that this is not necessarily the same as the complex-conjugate symmetry of the complex numbers!)

You can immediately work out the formulas for this algebraic conjugation operation on the fundamental quadratic irrationalities τ_d: if d is not congruent to 1 modulo 4, then $\tau_d = \sqrt{d}$, so obviously $\tau_d' = -\tau_d$, while if d is congruent to 1 modulo 4, then $\tau_d = \frac{1}{2}(1+\sqrt{d})$ and $\tau_d' = \frac{1}{2}(1-\sqrt{d}) = 1 - \tau_d$. This symmetry $\alpha \mapsto \alpha'$ respects all algebraic formulas. For example, to work out the algebraic conjugate of a polynomial expression like $\alpha\beta + 2\gamma^2$, where α, β, and γ are numbers in R_d, you just replace each individual number by its algebraic conjugate, obtaining the expression $\alpha'\beta' + 2\gamma'^2$.

The most telling integer quantity attached to a number $\alpha = x + y\tau_d$ in R_d is its *norm* $N(\alpha)$, which is defined to be the product $\alpha\alpha'$. This equals $x^2 - dy^2$ when $\tau_d = \sqrt{d}$ and $x^2 + xy - \frac{1}{4}(d-1)y^2$ when $\tau_d = \frac{1}{2}(1+\sqrt{d})$. The norm turns out to be *multiplicative*, meaning that $N(\alpha\beta) = N(\alpha)N(\beta)$, as you can directly check by multiplying out the formula for the norm of each factor and comparing with the norm of the product. This gives us a useful tactic for trying to factorize algebraic numbers in R_d, and offers criteria for determining whether a number α in R_d is a unit, and whether it is prime in R_d. In fact, an element $\alpha \in R_d$ is a unit if and only if $N(\alpha) = \alpha\alpha' = \pm 1$; in other words, the units are given by the integral solutions to the equations

$$X^2 - dY^2 = \pm 1 \tag{11}$$

or

$$X^2 + XY - \tfrac{1}{4}(d-1)Y^2 = \pm 1 \tag{12}$$

following the two cases. Here is the proof of this. If $\alpha = x + y\tau_d$ is a unit in R_d, then its reciprocal, $\beta = 1/\alpha$, must also be in R_d, and, of course, we have $\alpha\beta = 1$. Applying the norm to both sides of this equation and using the multiplicative property discussed above, we see that $N(\alpha)$ and $N(\beta)$ are reciprocal ordinary integers. Therefore, they are either both equal to $+1$ or both equal to -1. This shows that (x, y) is a solution to whichever of equation (11) or (12) is appropriate. In the other direction, if $N(\alpha) = \alpha\alpha' = \pm 1$, then the reciprocal of α is simply $\pm\alpha'$. This is in R_d so α is indeed a unit in R_d.

These homogeneous quadratic forms, the left-hand sides of equations (11) and (12) (which generalize formulas (3) and (9)), play an important role; let us refer to whichever of them is relevant to R_d as the *fundamental quadratic form* for R_d, and to its discriminant D as the *fundamental discriminant*. (D is equal to d if d is congruent to 1 modulo 4 and to $4d$ otherwise.) When d is negative there are only finitely many units (if $d < -3$ the only ones are ± 1) but when d is positive,

so that R_d consists entirely of real numbers, there are infinitely many. The ones that are greater than 1 are powers of a smallest such unit, ε_d, and this is called the *fundamental unit*.

For example, when $d = 2$ the fundamental unit, ε_2, is $1 + \sqrt{2}$, and when $d = 5$ it is the golden mean, $\varepsilon_5 = \frac{1}{2}(1 + \sqrt{5})$. Since any power of a unit is again a unit, we immediately have a machine for producing infinitely many units from any single one. For example, taking powers of the golden mean, we get

$$\varepsilon_5 = \tfrac{1}{2}(1+\sqrt{5}), \qquad \varepsilon_5^2 = \tfrac{1}{2}(3+\sqrt{5}),$$
$$\varepsilon_5^3 = 2+\sqrt{5}, \qquad \varepsilon_5^4 = \tfrac{1}{2}(7+3\sqrt{5}),$$
$$\varepsilon_5^5 = \tfrac{1}{2}(11+5\sqrt{5}),$$

all of which are units in R_5. The study of these fundamental units was already under way in the twelfth century in India, but in general their detailed behavior as d varies still holds mysteries for us today. For example, there is a deep theorem of Hua (1942) that tells us that $\varepsilon_d < (4e^2 d)^{\sqrt{d}}$ (for a proof of it along with a historical discussion of such estimates, see chapters 3 and 8 in Narkiewicz (1973)). There are examples of d that come close to attaining that bound, but we still do not know whether or not there is a positive number η and an infinity of square-free d for which $\varepsilon_d > d^{d^\eta}$. (The answer to this question would be yes if, for example, there were an infinity of R_d satisfying the unique factorization property! This follows from a famous theorem of Brauer (1947) and Siegel (1935); for a proof of the Brauer–Siegel theorem, see theorem 8.2 of chapter 8 in Narkiewicz (1973) or Lang (1970).)

6 Binary Quadratic Forms and the Unique Factorization Property

The principle of unique factorization is an all-important fact for the ring of ordinary integers $\mathbb{Z}$. The question of whether this principle does or does not hold for a given ring R_d is central to the algebraic number theory. There are helpful, analyzable, *obstructions* to the validity of unique factorization in R_d. These obstructions, in turn, connect with profound arithmetic issues, and have become the focus of important study in their own right. One such mode of expressing the obstruction to unique factorization is already prominent in Gauss's *Disquisitiones Arithmeticae* (1801), in which much of the basic theory of R_d was already laid down.

This "obstruction" has to do with how many "essentially different" binary quadratic forms $aX^2 + bXY + cY^2$ there are with discriminant equal to the fundamental discriminant D of R_d. (Recall that the discriminant of $aX^2 + bXY + cY^2$ is $b^2 - 4ac$, and that D equals $4d$ unless $d \equiv 1 \mod 4$, in which case it equals d.)

In order to define a binary quadratic form $aX^2 + bXY + cY^2$ of discriminant D, what you need to provide is simply a triplet of coefficients (a, b, c) such that $b^2 - 4ac = D$. Given such a form, one can use it to define other ones. For example, if we make a small linear change of the variables, replacing X by $X - Y$ and keeping Y fixed, then we get $a(X-Y)^2 + b(X-Y)Y + cY^2$, which simplifies to $aX^2 + (b-2a)XY + (c-b+a)Y^2$. That is, we get a new binary quadratic form whose triplet of coefficients is $(a, b-2a, c-b+a)$, and which (as can easily be checked) has the same discriminant D. We can "reverse" this change by replacing X by $X + Y$ and keeping Y fixed. If we do this reversal and perform the corresponding simplification then we get back our original binary quadratic form. Because of this reversibility, these two quadratic forms take exactly the same set of integer values as X and Y vary: it is therefore reasonable to think of them as *equivalent*.

More generally, then, one says that two binary quadratic forms are equivalent if one can be turned into the other (or minus the other) by any "reversible" linear change of variables with integer coefficients. That is, one chooses integers r, s, u, v such that $rv - su = \pm 1$, replaces X and Y by the linear combinations $X' = rX + sY$, $Y' = uX + vY$, and simplifies the resulting expression to get a new triplet of coefficients. The condition $rv - su = \pm 1$ guarantees that by a similar operation we can get back to our original binary quadratic form, and also that the new binary quadratic form has the same discriminant D as the old one. So when we talk of "essentially different" binary quadratic forms of discriminant D we mean that we cannot turn one into the other by this kind of change of variables.

Here is the surprising obstruction to unique factorization that Gauss discovered.

The unique factorization principle is valid in R_d if and only if every homogeneous quadratic form $aX^2 + bXY + cY^2$ with discriminant equal to the fundamental discriminant of R_d is equivalent to the fundamental quadratic form of R_d.

Furthermore, the collection of inequivalent quadratic forms whose discriminant is the fundamental discriminant of R_d expresses in concrete terms the degree to which R_d "enjoys unique factorization."

If you have never seen this theory of binary quadratic forms before, try your hand at working with quadratic forms in the case where $D = -23$. The idea is to start with some particular quadratic form $aX^2 + bXY + cY^2$ of your choice with discriminant $D = b^2 - 4ac = -23$. Then, using a sequence of carefully chosen linear changes of variables you reduce the size of the coefficients a, b, and c until you can go no further. Eventually you should end up with one of the two (inequivalent) quadratic forms that there are with discriminant -23: the fundamental form $X^2 + XY + 6Y^2$, or the form $2X^2 + XY + 3Y^2$. For example, can you see that the binary quadratic form $X^2 + 3XY + 8Y^2$ is equivalent to $X^2 + XY + 6Y^2$?

This type of exercise offers a small hint of the role that the *geometry of numbers* will play in the eventual theory. As you might expect from the venerability of these ideas, elegant streamlined methods have been discovered for making such calculations. Nevertheless, it is an open secret that any working mathematician, contemporary or ancient, engaged in this subject or nearby subjects, has done a myriad of straightforward simple hand computations along the lines of the above exercise.

If you try a few examples of this exercise, as I hope you do, here is one way of organizing your calculations. First, find a simple reversible linear change of variables to turn your form into an equivalent one with $a, b, c \geqslant 0$. (You may also have to multiply the whole form by -1.)

The cleanest way of writing down all binary quadratic forms given by triplets (a, b, c) of discriminant -23 is to list the triplets in increasing order of b, which will now be an odd positive integer. For each value of b you can then choose a and c in such a way that their product is $\frac{1}{4}(b^2 + 23)$. At this point the aim is to build up a repertoire of moves that tend to decrease b (which will keep a and c within bounds as well). A big clue, and aid, here is that for any pair of relatively prime integers x, y if you evaluate your quadratic form $aX^2 + bXY + cY^2$ at $(X, Y) = (x, y)$ to get the integer $a' = ax^2 + bxy + cy^2$, you can find, for appropriate b' and c', a quadratic form $a'X^2 + b'XY + c'Y^2$ equivalent to yours, with first coefficient a'. So, one tactic is to look for small integers represented by your quadratic form. Also the "example" linear change of variables $X \mapsto X - Y$, $Y \mapsto Y$ will lead you to be able to reduce the coefficient b to an integer smaller than $2a$. Can you check that $X^2 + XY + 6Y^2$ and $2X^2 + XY + 3Y^2$ are inequivalent?

Now, as we have just discussed, it follows from the general theory that R_{-23} does not have the unique factorization property. We can also see this directly. For example,

$$\tau_{-23} \cdot \tau'_{-23} = 2 \cdot 3,$$

and all four of the factors in this equation are irreducible in R_{-23}. To be a faithful companion, I should at this point give at least a hint at what connection there might be between this specific "failure of unique factorization" and the previous discussion. It may become a bit clearer in the next paragraph, but the underlying tension in the equation $\tau_{-23} \cdot \tau'_{-23} = 2 \cdot 3$ is that all the factors in our ring are prime: we are *missing* any elements in our ring R_{-23} that could factorize it further. We lack, for example, elements that play the role of the *greatest common divisor* of factors of this equation. The general theory regarding these matters (which we are not entering into here, but see EUCLID'S ALGORITHM [III.22]) tells us that what is missing is some element γ in R_{-23} that is both a linear combination of the numbers τ_{-23} and 2 (with coefficients in the ring R_{-23}) and also a common divisor of τ_{-23} and 2 in the ring R_{-23}, i.e., such that τ_{-23}/γ and $2/\gamma$ are both in R_{-23}. There is no such element, for its norm must divide $N(\tau_{-23}) = 6$ and $N(2) = 4$, and therefore be equal to 2, which can easily be shown to be impossible. But we are interested, rather, in the phenomenon that *inequivalence* of certain binary quadratic forms will indeed show this, so let us go on.

First, check that any linear combination

$$\alpha \cdot \tau_{-23} + \beta \cdot 2$$

with α, β elements of R_{-23} can also be written as $u \cdot \tau_{-23} + v \cdot 2$, where u and v are ordinary integers. Now compute the binary quadratic form given by systematically taking the norms of these linear combinations, and viewing these norms as functions of the integer coefficients u, v:

$$\begin{aligned} N(u \cdot \tau_{-23} + v \cdot 2) &= (\tau_{-23}u + 2v)(\tau'_{-23}u + 2v) \\ &= 6u^2 + 2uv + 4v^2. \end{aligned}$$

Viewing the u and the v as *variables*, and dubbing them U and V to emphasize their status as variables, we can say that the *norm quadratic form* obtained from the collection of linear combinations of τ_{-23} and 2 is

$$6U^2 + 2UV + 4V^2 = 2 \cdot (3U^2 + UV + 2V^2).$$

Now suppose that, contrary to fact, there *were* a common divisor, γ, as above; in particular, the multiples of γ in the ring R_{-23} would then be precisely the linear

combinations of the numbers τ_{-23} and 2. We would then have another way of describing those linear combinations; namely, for any pair of ordinary integers (u, v) there would be a pair of ordinary integers (r, s) such that

$$u \cdot \tau_{-23} + v \cdot 2 = \gamma \cdot (r\tau_{-23} + s) = r\gamma\tau_{-23} + s\gamma.$$

Taking norms, as above, we would get

$$\begin{aligned} N(\gamma \cdot (r\tau_{-23} + s)) &= N(r\gamma\tau_{-23} + s\gamma) \\ &= N(\gamma)(6r^2 + rs + s^2). \end{aligned}$$

Again, thinking of r and s as variables and renaming them R and S we would have the corresponding norm quadratic form:

$$N(\gamma) \cdot (6R^2 + RS + S^2) = 2 \cdot (6R^2 + RS + S^2).$$

Given the above facts—dependent, of course, on the contrary-to-fact hypothesis that there is a γ as above—the key idea is that there would be linear changes of variables from (U, V) to (R, S) and back that would establish an equivalence between the two quadratic forms $2 \cdot (3U^2 + UV + 2V^2)$ and $2 \cdot (6R^2 + RS + S^2)$. But these quadratic forms are not equivalent! Their inequivalence therefore shows that the putative γ does not exist and factorization in the ring R_{-23} is not unique.

7 Class Numbers and the Unique Factorization Property

In the previous section we saw that the collection of inequivalent quadratic forms of discriminant equal to the fundamental discriminant provides us with an obstruction to unique factorization. Somewhat later, a more articulated version of this obstruction arose, known as the *ideal class group* H_d of R_d. As its name implies, to describe this we must use the vocabulary of IDEALS [III.83 §2] and GROUPS [I.3 §2.1]. A subset I of R_d is an *ideal* if it has the following closure properties: if α belongs to I, so do $-\alpha$ and $\tau_d\alpha$, and if α and β belong to I, so does $\alpha + \beta$. (The first and third properties imply together that any integer combination of α and β belongs to I.) The basic example of such an ideal is the set of all multiples of some fixed, nonzero element γ of R_d, where by a *multiple* of γ we mean the product of γ and an element of R_d. We denote this set tersely as (γ), or, slightly more expressively, as $\gamma \cdot R_d$. An ideal of this sort, i.e., one that can be expressed as the set of all multiples of a single nonzero element γ, is called a *principal ideal*. For example, the ring R_d itself is an ideal (it consists, after all, of all linear combinations of 1 and τ_d) and is even a principal ideal: in our laconic terminology, it can be denoted $(1) = 1 \cdot R_d = R_d$. Strictly speaking, the singleton $\{0\}$ is also an ideal, but the ones that will interest us are the *nonzero ideals*.

As a direct counterpart to the obstruction principle involving binary quadratic forms that was described in the previous section, we have the following obstruction principle involving ideals.

The unique factorization principle is valid in R_d if and only if every ideal in R_d is principal.

Reflecting on this, you can get a sense of why the word "ideal" might have been chosen. Every principal ideal in R_d is of the form $\gamma \cdot R_d$ for some number γ in R_d (which is uniquely determined apart from multiplication by units), but sometimes there are more general ideals. These arise if you ever have two elements of R_d (think of τ_{-23} and 2, as in the previous section) such that the set of all their integer combinations *cannot* be expressed as the set of multiples of some fixed number γ in R_d. This phenomenon is a sign that we may be missing numbers in R_d that provide fine enough factorizations to make the arithmetic in R_d as smooth going as one might hope for. Just as a principal ideal $\gamma \cdot R_d$ corresponds to the number γ, ideals of this more general kind (think of the set of all integer combinations of τ_{-23} and 2) can be thought of as corresponding to "ideal numbers" that should, "by rights," be present in our ring, but happen not to be.

Once we think of ideals as standing for ideal numbers it makes some sense to try to multiply them: if I, J are two ideals in R_d, we let $I \cdot J$ denote the set of all finite sums of products $\alpha \cdot \beta$ in which α is in I and β is in J. The product of two principal ideals $(\gamma_1) \cdot (\gamma_2)$ is the principal ideal $(\gamma_1 \cdot \gamma_2)$ so, just as one would hope, multiplication of principal ideals corresponds to multiplication of the corresponding numbers. Multiplication of any ideal I by the ideal (1) leaves I unchanged: $(1) \cdot I = I$; we therefore refer to the ideal (1) as *the unit ideal*. With this new notion of *multiplication of ideals* we can now give the general definition of what it means for a prime number p to split or to remain prime in a ring R_d, the definition we promised in section 4.

The idea behind the definition is to use multiplication of ideals rather than of numbers. So if we are thinking about a prime p, the first thing we do is turn our attention to the principal ideal (p) in R_d. If this can be factorized as a product of two different ideals (*not* necessarily principal ideals, this is the whole point) in R_d, and if neither of these is the unit ideal $(1) = R_d$, then we say that p *splits* in R_d. If, on the other hand,

no factorization of the ideal (p) can be made without one of the factors being the ideal $(1) = R_d$, then we say that *p remains prime* in R_d. There is also a third important definition: if the principal ideal (p) can be expressed as the square of another ideal I, then we say that *p ramifies* in R_d. Continuing with the momentum of this definition, we may say that an ideal P is a *prime ideal* if P cannot be "factorized" as the product of two ideals neither of which is the unit ideal. This definition makes sense whether or not P is principal, so we are subtly shifting our attention from the multiplicative arithmetic of the numbers in R_d to the ideals.

By definition, two ideals are in the same *ideal class* if when you multiply each by an appropriate principal ideal you get the same ideal as a result. This is a natural EQUIVALENCE RELATION [I.2 §2.3] on ideals. It is also one that *respects products*, meaning that if I and J are two ideals, then the ideal class of their product $I \cdot J$ depends only on the ideal classes of I and J. (In other words, if I' is in the same ideal class as I and J' is in the same ideal class as J', then $I' \cdot J'$ is in the same ideal class as $I \cdot J$.) We can therefore say what we mean by *multiplication of ideal classes*: to multiply two classes, pick an ideal from each, multiply those, and take the ideal class of the resulting product. The set H_d of ideal classes of R_d, given this operation of multiplication, forms an Abelian group, in the sense that the multiplication law we have just defined is associative and commutative, and there are inverses. The identity element is the principal ideal R_d itself. This group H_d, the *ideal class group*, directly measures the extent to which the ideals of the ring R_d are principal: roughly speaking it is what you get if you take the multiplicative structure of all ideals and "divide out" by the principal ones.

As was mentioned in section 6, there is a close connection between ideal classes and binary quadratic forms. To begin to see this, take an ideal I of R_d and write it as the set of all integer combinations of two elements α, β of R_d. Then consider the norm function on the elements of I, that is,

$$\begin{aligned} N(x\alpha + y\beta) &= (x\alpha + y\beta)(x\alpha' + y\beta') \\ &= \alpha\alpha' x^2 + (\alpha\beta' + \alpha'\beta)xy + \beta\beta' y^2. \end{aligned}$$

This is a binary quadratic form in the variable coefficients x and y. If you start with a different choice of α, β that generate I you get a different form, but the two forms are scalar multiples of two forms with discriminant D that are equivalent to one another. Even better, the equivalence class of these forms depends only on the ideal class of I.

It can be shown that there are only a finite number of distinct ideal classes of R_d; that is, the ideal class group H_d is finite. The number of its elements is denoted h_d and called the *class number* of R_d. So, the obstruction to unique factorization of R_d is given by the nontriviality of the group H_d; equivalently, unique factorization holds for R_d if and only if its class number is 1. But whether or not H_d is trivial, its detailed group-theoretic structure is profoundly related to the arithmetic of R_d.

The class number enters into the generalizations of formulas (5) and (10) of section 1; that is, the *analytic formulas* we alluded to in that section. These formulas represent just the beginning of one of the ongoing chapters of our subject, and form a bridge between the world of discrete arithmetical issues and that of calculus, infinite series, and volumes of spaces, all of which can be attacked by the methods of COMPLEX ANALYSIS [I.3 §5.6]. Here is a sample of them.

(i) If $d > 0$ is a square-free integer and D is either d or $4d$ according to whether d is congruent to 1 modulo 4 or not, then

$$h_d \cdot \frac{\log \varepsilon_d}{\sqrt{D}} = \sum_{n \geqslant 0} \pm \frac{1}{n},$$

where the integers n run through those that are relatively prime to D and the signs $\pm$ are chosen in a way that depends only on the residue class of n modulo D.

(ii) If $d < 0$ we have a somewhat simpler formula: there is no fundamental unit ε_d in R_d to contend with, but when $d = -1$ or -3, there are more roots of unity than merely ± 1. If w_d denotes the number of roots of unity in R_d, then $w_{-1} = 4$, $w_{-3} = 6$ and otherwise $w_d = 2$, and then one has a formula of the following type:

$$\frac{h_d}{w_d \sqrt{D}} = \sum_{n \geqslant 0} \pm \frac{1}{n}.$$

As d tends to $-\infty$ the class number h_d tends to infinity.

We have effective lower bounds for the growth of h_d but these lower bounds are probably still far from the actual growth (cf. Goldfeld 1985). The effective lower bounds that are known are exceedingly weak. They follow, however, from beautiful work of Goldfield, and of Gross and Zagier: for every real number $r < 1$

there is a computable constant $C(r)$ such that $h_d > C(r)\log|D|^r$. Here is a sample:

$$h_d > \frac{1}{55}\prod_{p|D}\left(1-\frac{2\sqrt{p}}{p+1}\right)\cdot\log|D|$$

if $(D, 5077) = 1$.

It is a striking lacuna in our theory that, even today, nobody knows how to prove that there are infinitely many values of $d > 0$ for which R_d enjoys the unique factorization property—particularly since we expect that more than three quarters of them do! Our expectations are even more precise than that, thanks to Henri Cohen and Hendrik Lenstra, who make use of certain probabilistic expectations (now known as the *Cohen–Lenstra heuristics*) to conjecture that the density of positive fundamental discriminants of class number 1 among all positive fundamental discriminants is $0.75446\ldots$.

8 The Elliptic Modular Function and the Unique Factorization Property

A different obstruction to unique factorization in R_d is available when d is negative. Now R_d may be thought of as a lattice in the complex plane (see figure 3), which makes a wonderful tool available for us: the classical *elliptic modular function* of KLEIN [VI.57],

$$j(z) = e^{-2\pi iz} + 744 + 196\,884\,e^{2\pi iz} + 21\,493\,760\,e^{4\pi iz} + 864\,299\,970\,e^{6\pi iz} + \cdots. \tag{13}$$

This function, also colloquially referred to as the "j-function," converges for complex numbers $z = x + iy$ with $y > 0$. If $z = x + iy$ and $z' = x' + iy'$ are two such complex numbers, then $j(z) = j(z')$ if and only if the lattice generated by z and 1 in the complex plane is the same as the lattice generated by z' and 1 (or, equivalently, $z' = (az + b)/(cz + d)$, where a, b, c, and d are ordinary integers such that $ad - bc = 1$). We can paraphrase this by saying that the value $j(z)$ depends only on, and characterizes, the lattice generated by z and 1.

It turns out (by a theorem of Schneider) that if an algebraic number $\alpha = x + iy$ with $y > 0$ has the property that $j(\alpha)$ is also algebraic, then α is a (complex) quadratic irrationality; and the converse is also true. In particular, since $\alpha = \tau_d$ is such a complex quadratic irrationality when d is negative, the value, $j(\tau_d)$, of the j-function on τ_d is an algebraic number—in fact, an algebraic integer. This will be of some importance for our story. First, since the ring R_d as situated in the complex plane is simply the lattice generated by τ_d and 1, it follows from the previous paragraph that this value $j(\tau_d)$ will be the same if we replace τ_d by *any* element α of R_d, as long as the lattice generated by α and 1 is the entire ring R_d. More importantly, $j(\tau_d)$ is an algebraic integer of degree roughly comparable with the class number of R_d. In particular, it is an ordinary integer if and only if the ring R_d has the unique factorization property. (This result is one of the great applications of a classical theory known as *complex multiplication*.) In brief, here is yet another answer to the question of when the unique factorization principle holds for R_d when d is negative: if $j(\tau_d)$ is an ordinary integer, the answer is *yes*; otherwise it is *no*.

The search for the full list of negative values of d for which R_d has the unique factorization property makes a marvelous tale: there are precisely nine values of d for which it occurs (see below), but for over two decades number theorists, while knowing these nine, could prove only that there were no more than ten. The history of how the *nonexistence* of a possible tenth value of d was established, and reestablished, is one of the thrilling chapters in our subject. K. Heegner, in an article published in 1934, provided what he claimed was a proof of the nonexistence of the possible *tenth value of d*. However, Heegner's proof was framed in somewhat unfamiliar language and was not understood by the mathematicians of the time. His paper and his purported proof were largely forgotten until the late 1960s, when the nonexistence of the tenth field was established (to the mathematical community's satisfaction) by Stark (1967) and independently, via a different method, by Baker (1971). It was only then that mathematicians took a second and closer look at Heegner's original article and discovered that he had indeed proven exactly what he claimed. Moreover, his proof offered an elegant direct conceptual road to an understanding of the underlying issue.

Here are the nine values of d:

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

And here are the corresponding nine values of $j(\tau_d)$:

$$j(\tau_d) = 2^6 3^3,\ 2^6 5^3,\ 0,\ -3^3 5^3,\ -2^{15},\ -2^{15} 3^3,\ -2^{18} 3^3 5^3,\ -2^{15} 3^3 5^3 11^3,\ -2^{18} 3^3 5^3 23^3 29^3.$$

PUP: I can confirm that the fact that the second number is greater than the first in the sequence is OK here.

As Stark once pointed out, if, for some of these values of d, you simply "plug" τ_d into the power series expansion for j, you get rather surprising formulas. For

example, when $d = -163$, then

$$e^{-2\pi i\tau_d} = -e^{\pi\sqrt{163}}$$

is the first term of the power series for $j(\tau_{-163})$ (see formula (13)). Since $j(\tau_{-163}) = -2^{18}3^35^323^329^3$ and since all the terms $e^{2\pi n\tau_d}$ ($n > 0$) that appear in the power series for the j-function are relatively small, we find that $e^{\pi\sqrt{163}}$ is incredibly close to an integer. Indeed, it is $2^{18}3^35^323^329^3 + 744 + \cdots$, which works out as $262\,537\,412\,640\,768\,744 - \epsilon$, where the error term ϵ is less than 7.5×10^{-13}.

9 Representations of Prime Numbers by Binary Quadratic Forms

More often than you might expect, it turns out to be possible to translate difficult and/or somewhat artificial problems about ordinary integers into natural and tractable problems about larger rings of algebraic integers. My favorite elementary example of this type is the theorem due to FERMAT [VI.12] that if a prime number p may be expressed as a sum of two squares, $p = a^2 + b^2$ with $0 < a \leqslant b$, then it has only one such expression. (For example, $1^2 + 10^2$ is the only way of expressing the prime number 101 as the sum of two squares.) Moreover, a prime number p can be expressed as a sum of two squares if and only if $p = 2$ or p is of the form $4k + 1$. (The "only if" part of this is easy to see: since any square is congruent either to 0 or to 1 mod 4, an odd integer that is a sum of two squares is necessarily congruent to 1 mod 4.) These statements about ordinary integers can be translated into basic statements about the ring of Gaussian integers. For if we write $a^2 + b^2 = (a + ib)(a - ib)$, with $i = \sqrt{-1}$, then we can view $a^2 + b^2$ as the norm of the (conjugate) elements $a \pm ib$ in the ring of Gaussian integers. So, if p is a prime number that admits an expression as a sum of squares, $p = a^2 + b^2$, it follows that each of the elements $a \pm ib$ has norm a prime integer. It is easy to deduce that p is itself a prime in the ring of Gaussian integers. Indeed, any factorization of $a \pm ib$ into a product of two Gaussian integers would have the property that the norms of the factors are ordinary integers which multiply out to be the prime p, and this severely limits their possibilities: one of them has to be a unit.

In other words, whenever $p = a^2 + b^2$, then

$$p = (a + ib)(a - ib)$$

is a factorization of the ordinary integer prime p into a product of two Gaussian integer primes. The uniqueness part of Fermat's theorem then follows from (in fact, it is readily seen to be equivalent to) the unique factorization property of the ring R_{-1} of Gaussian integers. That any prime number p of the form $4k + 1$ admits such an expression as a sum of two squares follows from the *splitting law* for primes p in the ring of Gaussian integers: an odd prime number p is a norm, and hence splits into the product of two distinct primes, in the ring of Gaussian integers if and only if p is congruent to 1 mod 4. This result is just the beginning of an immense chapter of arithmetic.

10 Splitting Laws and the Race between Residues and Nonresidues

The simple *splitting law* for ordinary prime integers p in the ring of Gaussian integers, which states that p splits if $p \equiv 1 \bmod 4$ and not if $p \equiv -1 \bmod 4$, invites us to ask how often each of these cases occurs (see figure 4). DIRICHLET [VI.36] proved a famous theorem that says that there are infinitely many primes in the arithmetic progression $c, m + c, 2m + c, \ldots$ if the integers m and c are relatively prime. A more precise version of his result gives a clear asymptotic answer to the question we have just asked: as x goes to infinity, the ratio of the number of primes less than x that split to the number that do not tends to 1. (See ANALYTIC NUMBER THEORY [IV.2 §4] for a further discussion of Dirichlet's theorem.)

For fun, one might ask a fussier question: which type of prime less than x is actually in greater abundance, the nonsplit primes or the split ones (see figure 4)? To put some perspective on this, let us widen our query: for q equal either to 4 or to an odd prime, let $A(x)$ be the number of primes $\ell < x$ that are quadratic residues modulo q and let $B(x)$ be the number of primes $\ell < x$ that are quadratic nonresidues modulo q. Let $D(x) = A(x) - B(x)$ be the difference; what does $D(x)$ look like?

For an absorbing account of the history and status of this problem, see the article "Prime number races" by Andrew Granville and Greg Martin in *American Mathematical Monthly*.

PUP: what do you think of this sentence? The full reference details are available so perhaps I should add this to the further reading of this article and reword here instead, but this is how the author would prefer this to be cited.

11 Algebraic Numbers and Algebraic Integers

Now that we have seen the algebraic integers $j(\tau_d)$ for negative values of d, and have touched on trigonometric sums, we have a few hints that, as with ordinary integers, the deep structure of these rings of quadratic integers may be better understood within a larger context

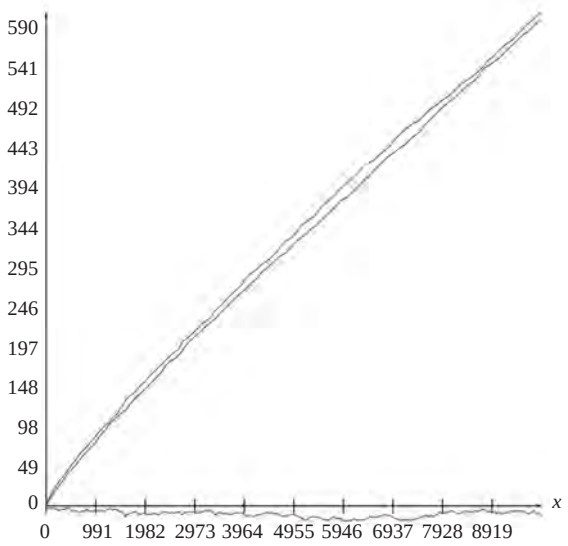

Figure 4 The higher of the two graphs in the figure represents the number of primes less than X that *remain prime* in the ring of Gaussian integers, and the lower represents the number of primes less than X that *split* in the ring of Gaussian integers. The third graph hovering around the x-axis represents the difference between the two numbers. We thank William Stein for this data.

of algebraic numbers. So now let us deal with algebraic numbers in full generality.

By a *monic* polynomial, we mean a polynomial of the form

$$P(X) = X^n + a_1X^{n-1} + \cdots + a_{n-1}X + a_n,$$

i.e., a polynomial of degree n such that the coefficient of X^n is 1. In general, the other coefficients are just assumed to be complex numbers. If $P(X) = X^n + a_1X^{n-1} + \cdots + a_{n-1}X + a_n$ is such a polynomial, and if Θ is a complex number such that $P(\Theta) = 0$, or, equivalently, if Θ satisfies the polynomial equation

$$\Theta^n + a_1\Theta^{n-1} + \cdots + a_{n-1}\Theta + a_n = 0,$$

we say that Θ is a *root* of the polynomial $P(X)$. THE FUNDAMENTAL THEOREM OF ALGEBRA [V.15], initially proved by Gauss, guarantees that any such polynomial of degree n factors into a product of n linear polynomials. That is,

$$P(X) = (X - \Theta_1)(X - \Theta_2)\cdots(X - \Theta_n)$$

for some complex numbers $\Theta_1, \Theta_2, \ldots, \Theta_n$ that are in fact precisely the roots of the polynomial $P(X)$.

If Θ is a root of such a polynomial $P(X) = X^n + a_1X^{n-1} + \cdots + a_{n-1}X + a_n$ and if in addition the coefficients a_i are rational numbers, then Θ is called an *algebraic number*. If the coefficients are not just rational but are in fact integers, then Θ is called an *algebraic integer*. So, for example, the square root of any rational number is an algebraic number and the square root of any "ordinary" integer is an algebraic integer. The same holds true for nth roots of ordinary integers, or of algebraic integers, for any natural number n. For an example of a different sort, we have already mentioned the theorem that the values of the j-function on complex quadratic irrational integers are algebraic integers. For a (random) particular case of that theorem, the complex number $j(\tau_{-23})$ is a root of the monic polynomial

$$X^3 + 3\,491\,750X^2 - 5\,151\,296\,875X + 12\,771\,880\,859\,375.$$

An exercise: show that any algebraic number can be expressed as an algebraic integer divided by an ordinary integer.

12 Presentation of Algebraic Numbers

In dealing with any mathematical concept, we confront, in one way or another, the dual problem of the various forms in which it comes to us when it arises in our work, and the various ways we can *present* it so as to deal with it effectively. We have already seen a bit of this at the outset of this article, in our discussion of quadratic surds, and we will continue to see it in our treatment of them below, where the various modes in which quadratic surds can be presented—as *radicals*, as *eventually recurrent continued fractions*, or as *trigonometric sums*—come together, all contributing to their unified theory.

This issue of presentation is all the more of a problem with algebraic numbers in general, which may come to us in a multitude of ways. For example, they can arise as the coordinates of points on specific algebraic varieties whose defining equations may not be easily available, or as special values of functions like the j-function. It is natural, then, to look for some uniform way of presenting algebraic numbers, and the history of the subject shows how much effort has been devoted to such a search. For example, consider the focus on iterated radical expressions, as in the famous formula for the solution to the general cubic equation $X^3 = bX + c$ given by

$$X = \left(\frac{c}{2} + \sqrt{\frac{c^2}{2} - \frac{b^3}{27}}\right)^{1/3} + \left(\frac{c}{2} - \sqrt{\frac{c^2}{2} - \frac{b^3}{27}}\right)^{1/3}, \quad (14)$$

or the corresponding general solution to the fourth-degree equation. These were major achievements of sixteenth-century Italian algebra, and they culminated in the proof that the general fifth-degree algebraic number could *not* be so expressed, which was a major achievement of the early nineteenth century (see THE INSOLUBILITY OF THE QUINTIC [V.24]). The challenge to give *some* analytic expression for such fifth-degree algebraic numbers was the source of a classic book by Klein, *The Icosahedron*, written in the late nineteenth century. Kronecker wrote that it was the "dream of his youth" (his *Jugendtraum*) to establish a uniform mode of presentation for a class of algebraic numbers that interested him, by expressing them as values of certain analytic functions.

13 Roots of Unity

A central role in the theory of algebraic numbers is played by the *roots of unity*, that is, the n complex solutions of the equation $X^n = 1$, or equivalently the n roots of the polynomial $X^n - 1$. If we let $\zeta_n = e^{2\pi i/n}$, then these roots are precisely ζ_n and its powers, so in particular they are algebraic integers. They give us the factorization

$$X^n - 1 = (X-1)(X-\zeta_n)(X-\zeta_n^2)\cdots(X-\zeta_n^{n-1}).$$

Now the powers of ζ_n form the vertices of a regular n-gon in the complex plane, centered at the origin. This has the following consequence, noticed by Gauss in his youth. It can be shown that compass and straightedge constructions allow us, in effect, to extract square roots, so whenever ζ_n can be given as an expression built out of just square roots and the usual arithmetical operations, we have, implicitly, a ruler-and-compass construction of the regular n-gon, and conversely.

To get some idea of why square roots are so closely connected with these constructions, consider this. If we have given ourselves a *unit measure*, which we can view as the distance between the numbers 0 and 1 in the (complex) plane, and if we have already constructed, by whatever device, a specific point, x say, between 0 and 1 on the horizontal axis of the plane, we can first "construct" $x/2$ by straightedge and compass, and then go on to form a right-angled triangle with hypotenuse of length $1 + x/2$ and one of its other sides of length $1 - x/2$ (again using a straightedge and compass). The Pythagorean theorem gives us that the third side of that triangle is of length $\sqrt{x}$. If one follows this line of thought (but adapts it to deal with complex quantities as well as the real number x as in the example we have just discussed), then one can see that the equations

$$\begin{aligned}
\zeta_3 &= \tfrac{1}{2}(1 + i\sqrt{3}),\\
\zeta_4 &= \sqrt{i},\\
\zeta_5 &= \tfrac{1}{4}(\sqrt{5}-1) + i\tfrac{1}{8}\Big(\sqrt{5+\sqrt{5}}\Big),\\
\zeta_6 &= -\tfrac{1}{2}(1 + i\sqrt{3})
\end{aligned}$$

provide (implicit) constructions of the equilateral triangle, the square, the regular pentagon, and the regular hexagon, respectively. By contrast, ζ_7 cannot be expressed solely in terms of the arithmetical operations and square roots (it is the root of a quadratic equation with coefficients that are rational expressions in the roots of the irreducible *cubic* polynomial $X^3 - \frac{7}{3}X + \frac{7}{27}$), which already suggests that the regular heptagon might fail to be constructible by the standard classical means—and indeed it does fail without some act of "angle trisection." (In principle, though, the reader can work out an expression for ζ_7 in terms of square roots and cube roots by means of the information provided in the parenthetical phrase above, together with equation (14).)

Gauss showed that if $n > 2$ is a prime number then the regular n-gon is classically constructible if and only if n is a *Fermat prime*, that is, a prime number of the form $2^{2^a} + 1$. So, for example, the 11-gon and 13-gon are not constructible by classical means, but since ζ_{17} is expressible as nested rational expressions of square roots, the 17-gon is, famously, constructible.

So, not all roots of unity can be expressed as iterated rational expressions of square roots. However, this inhospitability is not mutual, since all square roots of integers can be expressed as integer combinations of roots of unity. More mysteriously, the elusive fundamental units ε_d (for d positive), for which there is no known formula, are intimately related to a unit c_d in R_d which is an explicit rational expression of roots of unity. (See below: it is called a *circular unit*.) This satisfies the elegant formula

$$c_d = \varepsilon_d^{h_d}, \tag{15}$$

which establishes yet another explicit test of unique factorization: the equality $c_d = \varepsilon_d$ is a "litmus" requirement for the unique factorization principle to hold in R_d.

To give the flavor of the formulas involved, let p be an odd prime number and let a be an integer not divisible by p. Then define $\sigma_p(a)$ to be $+1$ if a is a *quadratic residue modulo* p, that is, if a is congruent to

the square of an integer modulo p, and -1 if not. The simple trigonometric sums of (1) and (6) generalize to *quadratic Gauss sums*:

$$\pm \mathrm{i}^{(p-1)/2}\sqrt{p} = \zeta_p + \sigma_p(2)\zeta_p^2 + \sigma_p(3)\zeta_p^3 + \cdots + \sigma_p(p-2)\zeta_p^{p-2} + \sigma_p(p-1)\zeta_p^{p-1}. \tag{16}$$

This formula is not too hard to prove, apart from determining which sign is correct in the initial $\pm$, but after considerable efforts Gauss managed to work this out too. To see the connection between, say, formula (6) and (16) note that when $p = 5$, the left-hand side of (16) is $\sqrt{5}$ and the right-hand side is

$$\zeta_5 + -\zeta_5^2 - \zeta_5^{-2} + \zeta_5^{-1} = 2\cos\tfrac{2}{5}\pi - 2\cos\tfrac{4}{5}\pi.$$

As for the circular unit c_p, it is defined to be

$$\prod_{a=1}^{(p-1)/2} (\zeta_p^a - \zeta_p^{-a})^{\sigma_p(a)} = \prod_{a=1}^{(p-1)/2} \sin(\pi a/p)^{\sigma_p(a)},$$

and this leads to further formulas. For example, when $p = 5$, we have $\varepsilon_p = \tau_5 = \frac{1}{2}(1+\sqrt{5})$, and since $h_5 = 1$, formula (6) for $p = 5$ tells us that

$$\frac{1+\sqrt{5}}{2} = \frac{\zeta_5 - \zeta_5^{-1}}{\zeta_5^2 - \zeta_5^{-2}} = \frac{\sin\frac{1}{5}\pi}{\sin\frac{2}{5}\pi}.$$

14 The Degree of an Algebraic Number

If Θ is an algebraic integer that is also a rational number, then Θ is an "ordinary" integer. Here is the proof of this fact. If Θ is a rational number, then we may write $\Theta = C/D$ as a fraction in lowest terms. If Θ is also an algebraic integer, then it is the root of a monic polynomial with rational integer coefficients, $\Theta^n + a_1\Theta^{n-1} + \cdots + a_n$, so we have an equation

$$(C/D)^n + a_1(C/D)^{n-1} + \cdots + a_{n-1}(C/D) + a_n = 0.$$

Multiplying through by D^n we get

$$C^n + a_1C^{n-1}D + \cdots + a_{n-1}CD^{n-1} + a_nD^n = 0,$$

where all terms are (ordinary) integers, and all but the first one is divisible by D. If $D > 1$ then it has some prime factor p, so all terms apart from the first are also divisible by p. Since the terms add up to zero, it follows that p divides C^n, which implies that p divides C, which contradicts the assertion that the fraction C/D is in its lowest terms. This in turn contradicts the hypothesis that Θ can be expressed as a ratio of whole numbers in the first place. As the reader may like to verify, this fact implies the result attributed to Theaetetus above, that $\sqrt{A}$ is irrational if and only if A is not a perfect square.

The *degree* of an algebraic number Θ is defined to be the smallest degree, n, of any polynomial relation $\Theta^n + a_1\Theta^{n-1} + \cdots + a_{n-1}\Theta + a_n = 0$ that Θ satisfies, where the coefficients a_i are rational numbers. The corresponding polynomial, $P(X) = X^n + a_1X^{n-1} + \cdots + a_{n-1}X + a_n$ is unique, since if there were two of them then their difference would be of smaller degree and would also have Θ as a root. (One could make it monic by dividing it through by the leading coefficient.) Let us call $P(X)$ the *minimal polynomial* of Θ. The minimal polynomial is *irreducible* over the field of rational numbers: that is, it cannot be factored as a product of two polynomials, each of smaller degree and having rational numbers as coefficients. (If it could, then it would not be of minimal degree, since one of its factors would have Θ as a root.) The minimal polynomial $P(X)$ of Θ is a factor of any monic polynomial $G(X)$ with rational coefficients that has Θ as root. (The greatest common divisor of P and G is another monic polynomial with rational coefficients that has Θ as a root, so it cannot be of degree smaller than that of P and it must therefore *be* P.) The minimal polynomial $P(X)$ of Θ has distinct roots. (If $P(X)$ had multiple roots, then a little elementary calculus shows that it would share a nontrivial factor with its derivative, $P'(X)$. Since the derivative is of lower degree than $P(X)$ and again has rational coefficients, the greatest common divisor of P and P' would provide a nontrivial factorization of $P(X)$, contradicting its irreducibility.)

A fundamental result due to Gauss is that the nth root of unity $\zeta_n = \mathrm{e}^{2\pi\mathrm{i}/n}$ is an algebraic integer of degree precisely $\phi(n)$, where ϕ is Euler's ϕ-function. For example, if p is prime, the minimal polynomial of ζ_p is

$$\frac{X^p - 1}{X - 1} = X^{p-1} + X^{p-2} + \cdots + X + 1,$$

which is of degree $\phi(p) = p - 1$.

15 Algebraic Numbers as Ciphers Determined by Their Minimal Polynomials

We have expressly insisted that our algebraic numbers are complex numbers (of a certain sort). But another possible attitude toward an algebraic number, Θ, an attitude at times promoted by Kronecker, among others, is to deal with Θ as an unknown satisfying only the algebraic relations implied by the fact that it is a root of its (unique monic) minimal polynomial with rational coefficients. For example, if the minimal polynomial of Θ is $P(X) = X^3 - X - 1$, then, according to this view,

Θ is just an algebraic symbol that comes with the rule that any occurrence of Θ^3 may be replaced by $\Theta + 1$ (rather as the complex number i can be regarded as a symbol with the property that i^2 may be replaced by -1). Any root of the minimal polynomial of Θ satisfies all the same polynomial relations with rational coefficients that Θ satisfies; these roots are called *conjugates* of Θ. If Θ is an algebraic number of degree n, then Θ has n *distinct* conjugates, all of them again, of course, algebraic numbers.

16 A Few Remarks about the Theory of Polynomials

Central to the theory of polynomials in one variable—and, therefore, particularly to the theory of algebraic numbers—is the general relationship that *roots* have to *coefficients*:

$$\prod_{i=1}^{n}(X - T_i) = X^n + \sum_{j=0}^{n-1}(-1)^j A_j(T_1, T_2, \dots, T_n)X^{n-j}.$$

The polynomial $A_j(T_1, T_2, \dots, T_n)$ is homogeneous of degree j (this means that every monomial in it has total degree j), has integer coefficients, and is symmetric in (i.e., unchanged by any permutation of) the variables $T_1, T_2, \dots, T_n$.

The constant term is the product of the roots:

$$A_n(T_1, T_2, \dots, T_n) = T_1 \cdot T_2 \cdot \cdots \cdot T_n,$$

which is known as the *norm* form. The coefficient of X^{n-1} is the sum of the roots:

$$A_1(T_1, T_2, \dots, T_n) = T_1 + T_2 + \cdots + T_n,$$

and this is the *trace* form.

When $n = 2$ the norm and trace are all the symmetric polynomials in the list. For $n = 3$, beyond the norm and trace we also have the symmetric polynomial of degree two:

$$\begin{aligned} A_2(T_1, T_2, T_3) &= T_1T_2 + T_2T_3 + T_3T_1 \\ &= \tfrac{1}{2}\{(T_1 + T_2 + T_3)^2 - (T_1^2 + T_2^2 + T_3^2)\}. \end{aligned}$$

It is of major importance to this theory, and more specifically to GALOIS THEORY [V.24], that the symmetry properties of the conjugate roots are nicely reflected in these symmetric polynomials. In particular, we have the fundamental result that *any* symmetric polynomial in $T_1, T_2, \dots, T_n$ with rational coefficients can be expressed as a polynomial with rational coefficients in the symmetric polynomials $A_j(T_1, T_2, \dots, T_n)$, and similarly with integral coefficients. For example, the equation above shows that $T_1^2 + T_2^2 + T_3^2$ can be expressed as

$$A_1(T_1, T_2, T_3)^2 - 2A_2(T_1, T_2, T_3).$$

17 Fields of Algebraic Numbers and Rings of Algebraic Integers

The inverse of a nonzero algebraic number is again an algebraic number; the sum, difference, and product of two algebraic numbers are algebraic numbers; the sum, difference, and product of two algebraic integers are algebraic integers. The neat proofs of these (latter) facts are a good demonstration of the power of linear algebra, and in particular of *Cramer's rule*. This states that any matrix with integer coefficients (and therefore also any linear transformation of a finite-dimensional vector space that preserves an integer lattice) satisfies a monic polynomial identity with integer coefficients.

To see just how useful this remark is for finding polynomial relations, and more specifically for showing that the collections of algebraic numbers and algebraic integers are closed under sums and products, try your hand at showing that $\sqrt{2} + \sqrt{3}$ is an algebraic integer. One way to do it is to search for the monic fourth-degree polynomial equation that it satisfies. But this is hardly a beautiful calculation! If, however, you are familiar with linear algebra, then a less painful route is to form the four-dimensional vector space over the rational numbers, generated by 1, $\sqrt{2}$, $\sqrt{3}$, and $\sqrt{6}$ (which are linearly independent when the scalars are rational). Multiplication by $\sqrt{2} + \sqrt{3}$ defines a linear transformation T of this vector space, and one can compute its characteristic polynomial P. The *Cayley–Hamilton theorem* says that $P(T) = 0$, and this translates into the statement that $\sqrt{2} + \sqrt{3}$ is a root of P.

These "closure properties" we have just discussed lead us to study, in complete generality, fields of algebraic numbers and rings of algebraic integers. A *number field* is a field that is generated (as a field) by finitely many algebraic numbers. A standard result tells us that any number field K can in fact be generated by a single carefully chosen algebraic number. The degree of this algebraic number equals the *degree* of K, which is defined to be the dimension of K when K is viewed as a vector space over the field $\mathbb{Q}$ of rational numbers. One of the main introductory observations of Galois theory is that if K is a number field of degree n, then there are exactly n distinct ring homomorphisms ("imbeddings") $\iota : K \to \mathbb{C}$ from K into the field of complex numbers. (This means that ι sends 1 to 1 and respects

the addition and multiplication laws within K. That is, $\iota(x+y) = \iota(x) + \iota(y)$ and $\iota(x \cdot y) = \iota(x) \cdot \iota(y)$.) From these imbeddings, we can construct some very useful rational-valued functions on K. For any element x in K, we form the n complex numbers $x_1, x_2, \ldots, x_n$ that are the images of x under the n different imbeddings of K into $\mathbb{C}$. We then let

$$a_j(x) = A_j(x_1, x_2, \ldots, x_n),$$

where $A_j(X_1, X_2, \ldots, X_n)$ is the jth symmetric polynomial of section 14 above. (Because the polynomials A_j are symmetric, we do not have to worry about the order of the images $x_1, x_2, \ldots, x_n$ in the above expression.) It is not immediately obvious that the values of a_j are rational numbers, but there is a theorem that tells us this.

If an algebraic number Θ in K generates K (as a field), then the rational numbers $a_j(\Theta)$ are the coefficients of its minimal polynomial; in general they are the coefficients of a power of its minimal polynomial. The most prominent of these functions are the multiplicative function $a_n(x) = x_1 \cdot x_2 \cdot \cdots \cdot x_n$, called the *norm* function, usually denoted $x \mapsto N_{K/\mathbb{Q}}(x)$, and the additive function $a_1(x) = x_1 + x_2 + \cdots + x_n$, called the *trace* function, usually denoted $x \mapsto \mathrm{trace}_{K/\mathbb{Q}}(x)$.

PUP: I can confirm that the use of '$a_n(x)$' and '$a_1(x)$' is OK.

The trace function can be used to define a fundamental symmetric bilinear form on the $\mathbb{Q}$-vector space K,

$$\langle x, y \rangle = \mathrm{trace}_{K/\mathbb{Q}}(x \cdot y),$$

which turns out to be nondegenerate. This nondegeneracy, together with the fact that if x, y are both algebraic integers, then $\langle x, y \rangle$ is an ordinary integer, can be used to show that the ring $\mathcal{O}(K)$ of *all* algebraic integers in K is finitely generated as an additive group. More specifically, there is a *basis* of algebraic integers in K, that is, a finite set $\{\Theta_1, \Theta_2, \ldots, \Theta_n\}$, such that any other algebraic integer in K can be expressed as an "ordinary" integer combination of the numbers Θ_i.

Let us summarize this structure. The number field K is a finite-dimensional vector space over $\mathbb{Q}$ and comes equipped with a nondegenerate bilinear symmetric form $(x, y) \mapsto \langle x, y \rangle$, and also with a lattice $\mathcal{O}(K) \subset K$. Moreover, the restriction of the bilinear form to $\mathcal{O}(K)$ takes on integral values.

The *discriminant* of K, denoted $D(K)$, is defined to be the DETERMINANT [III.15] of the matrix whose ij-entry is $\langle \Theta_i, \Theta_j \rangle$, for $\{\Theta_1, \Theta_2, \ldots, \Theta_n\}$ a basis of the lattice $\mathcal{O}(K)$; this determinant does not depend on the basis chosen.

The discriminant represents important information about the number field K. For one thing, there is a natural generalization to any number field of the notions of *splitting* and *ramification* that we discussed for quadratic fields, and the prime divisors p of $D(K)$ are precisely those prime numbers that ramify in the field extension K. By a theorem of MINKOWSKI [VI.64], the absolute value of the discriminant $D(K)$ of a number field K of degree n is always greater than

$$\left(\frac{\pi}{4}\right)^n \cdot \left(\frac{n^n}{n!}\right)^2.$$

This is greater than 1 unless K is the field of rational numbers. It follows that any nontrivial extension of the field of rational numbers has some prime that ramifies in it, a result that would be very hard to prove without the help of the algebraic structures we have just defined. This integer $D(K)$ really is quite a discriminating "tag" for our number field K, for, by a theorem of HERMITE [VI.47], given any integer D there are only finitely many different number fields with discriminant equal to D. (Not all integers can be discriminants: as is true for quadratic number fields, the integers D that are discriminants are either divisible by 4 or else congruent to 1 modulo 4.)

18 On the Size(s) of the Absolute Values of All Conjugates of an Algebraic Integer

As we have just seen, the coefficients of the minimal polynomial for an algebraic integer Θ are given by the ordinary integers $a_j(\Theta_1, \Theta_2, \ldots, \Theta_n)$, where the numbers Θ_i are all the conjugates of Θ. The sizes of all these coefficients must therefore all be less than some universal number M that depends only on the degree of Θ and the largest absolute value of any of its conjugates. As a consequence, given any n and any positive number B, there are only finitely many algebraic integers Θ of degree less than n such that the absolute values of Θ and its conjugates are all less than B. (This is because for any n and M there are only finitely many polynomials of degree less than or equal to n with the absolute values of all their integer coefficients at most M.) This finiteness result is the key to the following observation, due to Kronecker: if Θ is an algebraic number and if the absolute values of Θ and of all of its conjugates are equal to 1, then Θ is a root of unity. Indeed, all the powers of Θ have degree at most that of Θ, and they enjoy the same property: their absolute value, and that of all their conjugates, is equal to 1. Consequently, there are only finitely many such algebraic numbers, from which

it follows that there must be at least one coincidence of the form $\Theta^a = \Theta^b$ for different a and b. But this can happen only if Θ is a root of unity.

19 Weil Numbers

To follow this thread for just a bit, let us generalize the hypothesis of Kronecker's observation, and define a *Weil number*[3] of absolute value r to be a nonzero algebraic integer such that it and all of its conjugates have the same absolute value r. By the discussion in the previous section there are only finitely many distinct Weil numbers of given degree and absolute value. By Kronecker's theorem, which we have just described, the Weil numbers of absolute value 1 are precisely the roots of unity. Here are further basic facts that you might try to prove. First, the quadratic Weil numbers ω are precisely those quadratic algebraic integers such that $|\mathrm{trace}(\omega)| \leqslant 2\sqrt{|N(\omega)|} = 2\sqrt{|\omega\omega'|}$, where ω' is the (algebraic) conjugate of ω. Second, if p is prime then a quadratic Weil number ω of absolute value $\sqrt{p}$ is a prime element of the (unique) ring of quadratic integers R_d that contains ω, and therefore gives a prime factorization $\omega\omega' = \pm p$ of the integer p in that ring.

Weil numbers of absolute value $p^{\nu/2}$, where p is again a prime number and ν is a natural number, are extremely important in arithmetic: they hold the key to counting numbers of rational solutions of systems of polynomial equations over finite fields. For just one concrete example, the Gaussian integer $\omega = -1 + \mathrm{i}$ and its algebraic conjugate (which, in this instance, is also its complex conjugate) $\bar{\omega} = -1 - \mathrm{i}$ are Weil numbers (of absolute value 2) that control the number of solutions of the equation $y^2 - y = x^3 - x$ over all finite fields of size a power of 2. Specifically, the number of solutions of that equation over a field of order 2^ν is given by the formula

$$2^\nu - (-1-\mathrm{i})^\nu - (-1+\mathrm{i})^\nu$$

(which is an ordinary integer). This leads to another immense chapter of mathematics.

20 Epilogue

The single symmetry $\alpha \mapsto \alpha'$, the algebraic conjugation in the rings R_d that we have discussed, gave birth, thanks to ABEL [VI.33] and GALOIS [VI.41] in the beginning of the nineteenth century, to the rich study of (Galois) groups of symmetries of general number fields (see THE INSOLUBILITY OF THE QUINTIC [V.24]). This study continues with great intensity, since these Galois groups and their linear representations hold the key to a very detailed understanding of number fields. In its modern dress, algebraic number theory is closely connected with what is often called ARITHMETIC GEOMETRY [IV.5]. Kronecker's dream of getting explicit control of a wealth of algebraic number theoretic material by expressing algebraic numbers in terms of natural analytic functions has not yet been fully realized. Nevertheless, the scope of this dream (and, one might also add, the supply of natural analytic and algebraic functions) has expanded substantially: the full range of algebraic geometry and group representation theory is now being brought to bear on it. This is done, for example, by the *Langlands program*, which among other things works with objects known as *Shimura varieties*. On the one hand, these varieties have close connections with the theory of group representations and classical algebraic geometry, which greatly helps us to understand them. On the other hand, they are a rich source of concrete linear representations of Galois groups of number fields. This program, one of the glories of current mathematics, will, I expect, make a terrific chapter for a *Companion to Mathematics* to be written at the beginning of the next century.

Further reading

Basic Texts

First, I list three classics that require a minimum of background.

Gauss, C. F. 1986. *Disquisitiones Arithmeticae*, English edn. New York: Springer.
Davenport, H. 1992. *The Higher Arithmetic: An Introduction to the Theory of Numbers*. Cambridge: Cambridge University Press.
Hardy, G. H., and E. M. Wright. 1979. *Introduction to Number Theory*. Oxford: Oxford University Press.

At a more advanced level, the following are extraordinary expository books.

Borevich, Z. I., and I. R. Shafarevich. 1966. *Number Theory*. New York: Academic Press.
Cassels, J., and A. Fröhlich. 1967. *Algebraic Number Theory*. New York: Academic Press.
Cohen, H. 1993. *A Course in Computational Algebraic Number Theory*. New York: Springer.
Ireland, K., and M. Rosen. 1982. *A Classical Introduction to Modern Number Theory*, 2nd edn. New York: Springer.
Serre, J.-P. 1973. *A Course in Arithmetic*. New York: Springer.

3. This is a weaker condition than is usually required for Weil numbers but our deviation from standard usage should not be the cause of too much confusion.

Technical Articles and Books

Baker, A. 1971. Imaginary quadratic fields with class number 2. *Annals of Mathematics (2)* 94:139–52.
Brauer, R. 1950. On the Zeta-function of algebraic number fields. I. *American Journal of Mathematics* 69:243–50.
Brauer, R. 1950. On the Zeta-function of algebraic number fields. II. *American Journal of Mathematics* 72:739–46.
Goldfeld, D. 1985. Gauss's class number problem for imaginary quadratic fields. *Bulletin of the American Mathematical Society* 13:23–37.
Gross, B., and D. Zagier. 1986. Heegner points and derivatives of *L*-series. *Inventiones Mathematicae* 84:225–320.
Heegner, K. 1952. Diophantische Analysis und Modulfunktionen. *Mathematische Zeitschrift* 56:227–53.
Hua, L.-K. 1942. On the least solution of Pell's equation. *Bulletin of the American Mathematical Society* 48:731–35.
Lang, S. 1970. *Algebraic Number Theory*. Reading, MA: Addison-Wesley.
Narkiewicz, W. 1973. *Algebraic Numbers*. Warsaw: Polish Scientific Publishers.
Siegel, C. L. 1935. Über die Classenzahl quadratischer Zahlörper. *Acta Arithmetica* 1:83–86.
Stark, H. 1967. A complete determination of the complex quadratic fields of class-number one. *Michigan Mathematical Journal* 14:1–27.

IV.2 Analytic Number Theory

Andrew Granville

1 Introduction

What is number theory? One might have thought that it was simply the study of numbers, but that is too broad a definition, since numbers are almost ubiquitous in mathematics. To see what distinguishes number theory from the rest of mathematics, let us look at the equation $x^2 + y^2 = 15\,925$, and consider whether it has any solutions. One answer is that it certainly does: indeed, the solution set forms a circle of radius $\sqrt{15\,925}$ in the plane. However, a number theorist is interested in *integer* solutions, and now it is much less obvious whether any such solutions exist.

A useful first step in considering the above question is to notice that 15 925 is a multiple of 25: in fact, it is 25×637. Furthermore, the number 637 can be decomposed further: it is 49×13. That is, $15\,925 = 5^2 \times 7^2 \times 13$. This information helps us a lot, because if we can find integers a and b such that $a^2 + b^2 = 13$, then we can multiply them by $5 \times 7 = 35$ and we will have a solution to the original equation. Now we notice that $a = 2$ and $b = 3$ works, since $2^2 + 3^2 = 13$. Multiplying these numbers by 35, we obtain the solution $70^2 + 105^2 = 15\,925$ to the original equation.

As this simple example shows, it is often useful to decompose positive integers multiplicatively into components that cannot be broken down any further. These components are called *prime numbers*, and THE FUNDAMENTAL THEOREM OF ARITHMETIC [V.16] states that every positive integer can be written as a product of primes in exactly one way. That is, there is a one-to-one correspondence between positive integers and finite products of primes. In many situations we know what we need to know about a positive integer once we have decomposed it into its prime factors and understood those, just as we can understand a lot about molecules by studying the atoms of which they are composed. For example, it is known that the equation $x^2 + y^2 = n$ has an integer solution if and only if every prime of the form $4m+3$ occurs an even number of times in the prime factorization of n. (This tells us, for instance, that there are no integer solutions to the equation $x^2 + y^2 = 13\,475$, since $13\,475 = 5^2 \times 7^2 \times 11$, and 11 appears an odd number of times in this product.)

Once one begins the process of determining which integers are primes and which are not, it is soon apparent that there are many primes. However, as one goes further and further, the primes seem to consist of a smaller and smaller proportion of the positive integers. They also seem to come in a somewhat irregular pattern, which raises the question of whether there is any formula that describes all of them. Failing that, can one perhaps describe a large class of them? We can also ask whether there are infinitely many primes. If there are, can we quickly determine how many there are up to a given point? Or at least give a good estimate for this number? Finally, when one has spent long enough looking for primes, one cannot help but ask whether there is a quick way of recognizing them. This last question is discussed in COMPUTATIONAL NUMBER THEORY [IV.3]; the rest motivate the present article.

Now that we have discussed what marks number theory out from the rest of mathematics, we are ready to make a further distinction: between *algebraic* and *analytic* number theory. The main difference is that in algebraic number theory (which is the main topic of ALGEBRAIC NUMBERS [IV.1]) one typically considers questions with answers that are given by exact formulas, whereas in analytic number theory, the topic of this article, one looks for *good approximations*. For the sort of quantity that one estimates in analytic number theory, one does not expect an exact formula to exist, except perhaps one of a rather artificial and unilluminating kind. One of the best examples of such a