THE FUNDAMENTAL THEOREM OF ARITHMETIC

ANDREW GRANVILLE

1. Introduction

1.1. **The Fundamental Theorem.** The *positive integers* are the integers $1, 2, 3, \ldots$. The *prime numbers* are those integers larger than 1 that can be factored into two positive integers in exactly one way (not paying attention to order). Thus $2, 3, 5, 7, 11, \ldots$ are primes, whereas $1, 4, 6, 8, 9, 10, \ldots$ are not primes. These non-prime integers > 1 are called *composite numbers*: to see that 10 is composite note that we can *factor* it in two distinct ways, as 1×10 and as 2×5.

When one studies questions involving integers one quickly finds that it is useful to break integers down into their smallest component parts, that is to factor them into prime numbers. Thus 35 is 5×7, and 90 is $2 \times 3 \times 3 \times 5$, and so on; in fact, every positive integer can be factored in such a manner. A factorization into primes cannot be decomposed any further since none of the component primes can be factored again. From calculations it appears that there is only one way to factor a given integer, though this does not seem to be so easy to prove. However if true then it does give a solid foundation to any study of the positive integers, and so the result is considered to be the most fundamental in arithmetic:

The fundamental theorem of arithmetic. *Every integer > 1 may be factored as a product of primes in a unique way.*

It must be stressed that the primes involved in a factorization are not necessarily distinct (as in $12 = 2 \times 2 \times 3$), and that we consider the same primes written in two different orders as the same factorization (that is, $30 = 2 \times 3 \times 5$ and $5 \times 2 \times 3$ are considered to be the same factorization). The easiest "canonical" way to display n as a given product of primes is to write $n = p_1^{e_1} p_2^{e_2} \ldots p_k^{e_k}$ for primes $p_1 < p_2 < \cdots < p_k$ and positive integers $e_1, e_2, \ldots, e_k$.[1]

Many ancient authors were interested in *perfect* numbers (integers equal to the sum of their proper divisors, like 6 and 28) and pairs of *amicable* numbers (each equal to the sum of the other's proper divisors, like the pair 220 and 284), which meant that they

L'auteur est partiellement soutenu par une bourse du Conseil de recherches en sciences naturelles et en génie du Canada.

[1] Sometimes, though, it is convenient to allow some of the e_is to be zero.

needed to be able to determine the divisors of a given integer: In fact if $n = p_1^{n_1} p_2^{n_2} \dots p_k^{n_k}$ for primes $p_1 < p_2 < \cdots < p_k$ and positive integers $e_1, e_2, \dots, e_k$, then we can deduce from the fundamental theorem of arithmetic that the divisors m of n are the integers of the form $p_1^{m_1} p_2^{m_2} \dots p_k^{m_k}$ where m_j is an integer with $0 \leq m_j \leq n_j$.

Another consequence of the fundamental theorem of arithmetic is that we can easily determine the *greatest common divisor* of any two given integers m and n, for if $m = \prod_{i=1}^k p_i^{m_i}$ and $n = \prod_{i=1}^k p_i^{n_i}$ then their greatest common divisor, denoted by (m, n), equals $\prod_{i=1}^k p_i^{\min\{m_i, n_i\}}$ (note though that this is only really "easy" if we have the factorizations of m and n). What the ancient Greeks realized is that it is possible to determine the greatest common divisor of two non-negative integers *without* knowing their factorizations — the method is now called *the Euclidean algorithm*. We start with two integers $n \geq m > 0$ and then let $\ell = n - m$ so that ℓ is also a non-negative integer and a multiple of the greatest common divisor of m and n, for if $(m, n) = g$ with $m = gM$ and $n = gN$ then $\ell = g(N - M)$. Therefore (m, n) is a common divisor of both ℓ and m, and hence $(m, n) \leq (\ell, m)$. On the other hand, since $n = \ell + m$, the greatest common divisor of ℓ and m divides n and so $(\ell, m) \leq (m, n)$ by the analogous reasoning. Putting these two facts together implies that $(m, n) = (\ell, m)$, so that the greatest common divisor of m and n is equal to the greatest common divisor of two smaller integers, ℓ and m. The Euclidean algorithm consists of repeating this process finishing only when one of the integers is 0, and it must finish in a finite number of steps since there are only finitely many non-negative integers up to any given n. As an example we see that $(22, 8) = (14, 8) = (8, 6) = (6, 2) = (4, 2) = (2, 2) = (2, 0) = 2$; evidently this can be speeded up by writing $(n, m) = (m, r)$, where r is the least non-negative residue of $n \pmod m$, and therefore $(22, 8) = (8, 6) = (6, 2) = (2, 0) = 2$.

But there is more: We see that ℓ and m are both integral linear combinations of m and n; and indeed the next two integers in the Euclidean algorithm are integral linear combinations of ℓ and m, and thus of m and n. Continuing like this we deduce that the greatest common divisor of m and n is also an integral linear combination of m and n; that is, we have integers u and v for which

$$(m, n) = mu + nv.$$

For example $2 = 22 \times (-1) + 8 \times 3$.

This surprising observation allows us to give an elegant though unintuitive proof of the fact that if a prime p divides the product of two integers a and b then it divides at least one of them. For if p does not divide a then $(p, a) = 1$,[2] and therefore there exist

[2] Since (p, a) must be a divisor of p, and so either 1 or p, and yet it is not p as p does not divide a.

integers u and v for which $pu + av = 1$. Therefore $pbu + (ab)v = b$ and so p divides b since p divides both $p(bu)$ and $(ab)u$. We next deduce, by induction, that if a prime p divides a product of integers then it divides at least one of them.

Finally we are ready to prove that there is only one factorization of any given integer: if $p_1p_2\dots p_k = q_1q_2\dots q_\ell$ is the smallest counterexample (where two of the p_is, or two of the q_js, may be equal) then q_ℓ divides $p_1p_2\dots p_k$ so must divide one of them, say p_k, so we have a smaller counterexample $p_1p_2\dots p_{k-1} = q_1q_2\dots q_{\ell-1}$, a contradiction.

This collection of ideas has inspired many developments in number theory, algebra and beyond, as we will discuss.

1.2. **A confused history.** The key ideas in the fundamental theorem of arithmetic have probably been recognized by any society that thought deeply about mathematics, and it was the genius of mathematicians in ancient Greece (and possibly Mesopotamia), and then Egypt, Turkey, India, North Africa and beyond, to realize that such statements, arguably "self-evident", would be best justified by proofs deduced from even more transparent propositions. Most of these older mathematical cultures recognized the fact that integers can be factored into primes as an essential step in determining all of the divisors of a given integer (as we did above). In so doing they almost certainly must have assumed, perhaps unknowingly, that their given factorization of an integer is the only one; it was the genius of the young Gauss to realize that this fundamental observation needs proof and then becomes the cornerstone of the theory of numbers. This has subsequently been celebrated as some of the most agile and deft reasoning in the history of human thought.

Those parts of Euclid's *Elements* that survive from ancient times are among the earliest known mathematical texts. Much is remarkable about these books, and the effort therein to put mathematics on a sound axiomatic footing was not truly surpassed until about two thousand years later. We will perhaps never know which parts of these books were original to Euclid, though I believe that the succinct, irrefutable proofs given, indicate that Euclid must have been a leading participant in a sophisticated mathematical culture. When reading Euclid's work today we must be careful of (at least) two cultural issues:

- Euclid's objectives reflected the questions and thinking of his day, not ours, at a time when "publishing" was, by today's standards, unimaginably expensive. Hence what he chose to present cannot properly be judged by what we would choose to present today.
- The notation of those times was far less flexible than that of today, so that the astute reader necessarily had to deduce the full content of the statement of a theorem, or of a proof, from what was written, and could not necessarily learn all that was meant

from what was actually written.[3] Looking back it may seem unimaginable that the finest minds of that time could not recognize this limitation in their notation and do something about it, yet even in the Renaissance, Fermat and Descartes recognized this difficulty and deplored those who could not navigate it adroitly.

Euclid's number theory begins with the Euclidean algorithm, which gives him a notion of two integers being relatively prime. From this he deduces (in Book 7, proposition 30) that if p divides the product of two integers a and b then it divides at least one of them,[4] and deduces (in proposition 31) that every integer has a prime factor. Then much later, in Book 9, proposition 14, almost as an afterthought, he proves that a product of distinct prime numbers is not divisible by any other prime number, that is he proves the unique factorization theorem for squarefree numbers.

It is easy to deduce the fundamental theorem of arithmetic from these propositions in Euclid, and there can be little doubt that had he recognized this result as fundamental he would have proved it. Euclid was more interested in being able to list (with proof) all of the divisors of certain integers. For example, a *perfect number* is an integer which equals the sum of its proper divisors and Euclid observed (in Book 9, proposition 36) that $2^{p-1}q$ is a perfect number whenever $q = 2^p - 1$ is prime.

The oldest surviving text with a clear statement that every positive integer can be written as a finite product of prime numbers was given by al-Farisi in Persia around 1300. In his text on amicable pairs, he exhibited the pair $2^k pq, 2^k r$ whenever $p = 3 \cdot 2^{k-1} - 1, q = 3 \cdot 2^k - 1$ and $r = 9 \cdot 2^{2k-1} - 1$ are all prime, $k \geq 2$.

Even Renaissance mathematicians such as Euler and Legendre failed to note the importance of the uniqueness of factorization, and it was not until the streamlined beauty of Gauss's [5], where in article 16 we finally read

A composite number can be factored into prime factors in one and only one way,

where he fully credits Euclid for all of the essential ideas that go into this statement. See [3] and [8] for further discussion.

1.3. **Continued fractions.** We shall re-work Euclid's algorithm and its generalizations in various ways to highlight different ideas. Perhaps the most ancient is by determining the continued fraction of m/n, for positive integers m and n. For example if $m = 30, n = 13$ then in Euclid's algorithm we begin by noting that $13 \times 2 \leq 30 < 13 \times 3$

[3] For example, when Euclid proves the infinitude of primes (Book 9 proposition 20), he gives a proof by contradiction assuming that there are just three primes. The reader is evidently meant to infer that the same proof works no matter how large a finite number of primes we assume there to be.

[4] Actually he proves what is now known as Euclid's lemma: if d divides ab with $(d, a) = 1$ then d divides b.

so that we take $4 = 30 - 13 \times 2$. Thus we go from considering the fraction $30/13$ to considering the fraction $4/13$, which comes up as:

$$\frac{30}{13} = 2 + \frac{4}{13}.$$

Note that $2 = [30/13]$, where $[t]$ denotes the largest integer $\leq t$. In Euclid's algorithm we want the larger number first, so we consider the fraction $13/4$ instead of $4/13$; that is we invert the fraction under consideration:

$$\frac{30}{13} = 2 + \frac{1}{13/4}.$$

Now we repeat the above process: First we have $3 = [13/4]$, then $13/4 = 3 + 1/4$, and so we have the *continued fraction*

$$\frac{30}{13} = 2 + \frac{1}{3 + \frac{1}{4}}.$$

If Euclid's algorithm takes many steps for a particular pair m, n then the continued fraction will be long, and difficult to typeset, so we use the more convenient notation

$$30/13 = [2, 3, 4].$$

There is one ambiguity in this notation in that we could equally have written $[2, 3, 3, 1]$, but we make the choice never to end a finite continued fraction with a '1'.

One can create a continued fraction for any real number α: one has $\alpha = [a_0, a_1, \dots]$ where $a_0 = [\alpha]$ and $[a_1, a_2, \dots] = 1/(\alpha - a_0)$. Typically we write $p_n/q_n = [a_0, a_1, \dots, a_n]$, and one can show from the definition that $p_{2k}/q_{2k} \leq \alpha \leq p_{2k+1}/q_{2k+1}$ for all $k \geq 0$.

There is another, rather useful, way to represent continued fractions, in terms of 2-by-2 matrices, which was discovered surprisingly recently (in the 1940s). We begin by considering our pair as a point (m, n) in the plane, and then determine all of our pairs of integers as such points, via linear transformations of the corresponding point. Thus

$$\begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 30 \\ 13 \end{pmatrix} = \begin{pmatrix} 4 \\ 13 \end{pmatrix}$$

and

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 4 \\ 13 \end{pmatrix} = \begin{pmatrix} 13 \\ 4 \end{pmatrix},$$

which together yield

$$\begin{pmatrix} 0 & 1 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} 30 \\ 13 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 30 \\ 13 \end{pmatrix} = \begin{pmatrix} 13 \\ 4 \end{pmatrix}.$$

Multiplying through by the inverse of our 2-by-2 matrix yields

$$\begin{pmatrix} 30 \\ 13 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 13 \\ 4 \end{pmatrix};$$

and therefore

$$\begin{pmatrix} 30 \\ 13 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 3 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 4 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} g \\ 0 \end{pmatrix},$$

where $g = (30, 13) = 1$. In fact

$$\begin{pmatrix} 30 & 7 \\ 13 & 3 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 3 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 4 & 1 \\ 1 & 0 \end{pmatrix}$$

where $30/13 = [2, 3, 4]$ and $7/3 = [2, 3]$. Taking determinants of these matrices we see that $30 \cdot 3 - 13 \cdot 7 = -1$, that is we have the integral linear combination of 30 and 13 that gives 1.

For any real α this argument generalizes to give

$$\begin{pmatrix} p_n & p_{n-1} \\ q_n & q_{n-1} \end{pmatrix} = \begin{pmatrix} a_0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a_1 & 1 \\ 1 & 0 \end{pmatrix} \cdots \begin{pmatrix} a_n & 1 \\ 1 & 0 \end{pmatrix}$$

so that $p_n q_{n-1} - p_{n-1} q_n = (-1)^{n-1}$, and therefore $|\alpha - p_n/q_n| \leq |p_{n+1}/q_{n+1} - p_n/q_n| = 1/q_n q_{n+1} \leq 1/a_n q_n^2$.

All of the steps from this example generalize directly to any m/n provided $m, n \geq 0$. It is worth understanding the geometry involved in this representation. The points all belong to the upper right quadrant of the complex plane. We begin with a point on or to the right of the line $y = x$. The first step described, subtracting a suitable integer a, translates our original point horizontally, by an integer multiple of y, to the unique such point with the same y-value, but with x-value to the left of the line $y = x$ while remaining in the same quadrant. This step, which can be thought of as a copies of the basic step of size y to the left, is written in matrix form as $\begin{pmatrix} 1 & -a \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}^a$. The second step begins with a point to the left of the line $y = x$. We reflect this point in the line $y = x$ which creates a point with a smaller y-value; this step is written in matrix form as $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. With each such pair of steps we find a new point that is both lower and nearer to the origin than the point we started off with, and we keep on going until we reach the origin. The last point before we reach the origin will be on the x-axis with co-ordinates $(g, 0)$, where $g = \gcd(m, n)$. This interpretation, and the two basic transformations involved, will come in useful again later when we work with more complicated numbers.

1.4. **Square Roots.** We can use the fundamental theorem of arithmetic to show that if an integer n is the square of a rational number then it must be the square of an integer:[5] If $m = \prod_{i=1}^k p_i^{m_i}$ then $m^2 = \prod_{i=1}^k p_i^{2m_i}$, and so $n = \prod_{i=1}^k p_i^{n_i}$ with $p_1 < p_2 < \cdots < p_k$ is the square of an integer if and only if each n_i is even. Thus we ask whether we can write $n = \prod_{i=1}^k p_i^{n_i}$ as the square of a rational number when at least one of the exponents n_i is odd, say n_j. If that rational number is a/b or $-a/b$, we have $n = (a/b)^2$ so that $a^2 = nb^2$. If the exact power of p_j dividing a and b are a_j and b_j, respectively, then we deduce that the exact power of p_j dividing a^2 is $2a_j$ which is even, and the exact power of p_j dividing nb^2 is $n_j + 2b_j$ which is odd. This contradicts the fundamental theorem of arithmetic. Thus we have proved that $\sqrt{2}$ is irrational and, in fact, $\sqrt{n}$ is irrational where n is any squarefree positive integer.

The number $\sqrt{2}$ arises naturally in many contexts in mathematics (for instance as the hypotenuse of the right-angled triangle with smaller two sides both of length 1); hence it is of interest to understand the arithmetic of numbers of the form $a + b\sqrt{2}$ where a and b are integers. We can ask whether some analogy of the fundamental theorem of arithmetic holds for such numbers? When one attempts to again copy over our original proof, one reaches an unexpected barrier: For the usual integers we used the fact that there are only finitely many positive integers less than a given integer, and for polynomials that there are only finitely many possible degrees less than a given degree. In either case this comes from the fact that there is a smallest positive integer. Here we would need that there is a smallest positive number of the form $r + s\sqrt{2}$ with r and s integers, but this is not true! To exhibit this fact we use an elegant argument due to Dirichlet: For any given real number t we define $\{t\}$ to be the *fractional part* of t, in other words $\{t\} = t - [t]$. Note that $0 \le \{t\} < 1$ for every real number t. Now suppose that there is a smallest positive number of the form $r + s\sqrt{2}$ with r and s integers, and select integer N sufficiently large that $r + s\sqrt{2} > 1/N$. The numbers $0, \{\sqrt{2}\}, \{2\sqrt{2}\}, \{3\sqrt{2}\}, \dots, \{N\sqrt{2}\}$ all lie between 0 and 1, so two of them, say $\{i\sqrt{2}\}$ and $\{j\sqrt{2}\}$ with $0 \le i < j \le N$, must lie a distance no more than $1/N$ apart. We can write $i\sqrt{2} = r_i + \{i\sqrt{2}\}$ and $j\sqrt{2} = r_j + \{j\sqrt{2}\}$ for some integers r_i and r_j, so that for $a = r_j - r_i$ and $b = i - j$ we have

$$|a + b\sqrt{2}| = |\{i\sqrt{2}\} - \{j\sqrt{2}\}| \le \frac{1}{N} < r + s\sqrt{2},$$

so that either $a + b\sqrt{2}$ or $-a - b\sqrt{2}$ contradicts the minimality of $r + s\sqrt{2}$.

[5] This result is credited to the young Theaetetus in Plato's dialogue of that name, dating from around 390 B.C.

Euclid did recognize the importance of the fact that there is a smallest positive integer[6], but it was not until the nineteenth century that anyone found out how one might extend such ideas beyond this barrier.

2. Unique factorization in other domains?

2.1. **Polynomials.** One learns early in mathematics that once one finds *the* roots of a given polynomial, then one can completely and uniquely factor the polynomial. This statement is more subtle than it might seem at first sight for it presupposes that there is just one way to factor a polynomial (that is, it is impossible to find more than one way to factor a polynomial). We have to be careful with such a harmless looking statement for if we consider the very simple polynomial $x^2 - 1 = (x-1)(x+1)$, not in its usual context but rather working $\pmod m$ for various integers m,[7] then we see that this simple assumption fast breaks down since it has the additional factorizations $(x-3)(x+3) \pmod 8$, $(x-4)(x+4) \pmod{15}$, etc. Nonetheless, in the usual context, we do have the following fundamental theorem: Every polynomial with coefficients in $\mathbb{C}$ may be factored as a scalar times a product of monic,[8] linear polynomials in a unique way.

We can prove this in much the same way as we proved the fundamental theorem of arithmetic, by demonstrating that a suitably modified Euclidean algorithm works in this context. Here the greatest common divisor of two polynomials is the monic polynomial of highest degree which divides both of the two original polynomials. Thus if we begin the Euclidean algorithm with two polynomials f and g which have leading terms ax^d and Ax^D, respectively, where $d \geq D$ and a and A are non-zero, then we define $h = f - (a/A)x^{d-D}g$ and prove that $(f,g) = (g,h)$. As the degrees of g and h are smaller than those of f and g we see that this process will terminate in finitely many steps. This is a useable analogy to the Euclidean algorithm for integers, and the fundamental theorem in this context is then proved entirely analogously.

The same algorithm, suitably modified, also works for pairs of polynomials mod p where p is prime, though not modulo composites. The key issue being that we need to be able to invert the leading non-zero coefficient A above, which cannot necessarily be

[6] See, e.g., Book 7, proposition 31 in which he proves that every integer has a prime factor.

[7] Two polynomials f and g with integer coefficients are *congruent* $\pmod m$ if $f - g$ is m times a polynomial with integer coefficients.

[8] By *leading coefficient* we mean the coefficient of the power of x of highest degree in the polynomial. A polynomial is *monic* if its leading coefficient is 1. Therefore monic linear polynomials are those of the form $x - \alpha$ for some $\alpha \in \mathbb{C}$.

done if the ring of integers modulo m contains zero divisors (e.g. $4 \cdot 2 \equiv 0 \pmod 8$ and $5 \cdot 3 \equiv 0 \pmod{15}$).

2.2. **Where there is no unique factorization!** We have already seen that the polynomials mod 15, and other non-prime moduli, do not all have an unique factorization. This is not so surprising when we are working in situations in which there are zero divisors (in this case $3 \times 5 \equiv 0 \pmod{15}$), that is where there are non-zero integers r, s such that $rs \equiv 0 \pmod m$. So do we have unique factorization in domains in which there are no zero divisors?

The set of numbers $\{a + b\sqrt{-6} : \ a, b \in \mathbb{Z}\}$ is a *ring* but we have two factorizations of 6, namely $-1 \times \sqrt{-6} \times \sqrt{-6}$ and 2×3, where $\sqrt{-6}, 2$ and 3 are all *irreducible* in our ring; in other words they cannot be written as a product of two other numbers in the ring, neither of which is 1 or -1. To see that neither 2 nor 3 can be so written note that if an integer equals $(a + b\sqrt{-6})(c + d\sqrt{-6})$ then $ad + bc = 0$; that is there exist coprime integers r, s such that $a + b\sqrt{-6} = t(r + s\sqrt{-6})$ and $c + d\sqrt{-6} = u(r - s\sqrt{-6})$ for some integers t and u, so that our integer is $tu(r^2 + 6s^2)$ and thus evidently not 2 or 3 unless $r = \pm 1$ and $s = 0$ which does not give rise to a factorization. So the ring $\{a + b\sqrt{-6} : \ a, b \in \mathbb{Z}\}$ does not have unique factorization! If we are going to be able to study its arithmetic we are going to need a way around this deficiency.

2.3. **Proving Fermat's Last Theorem.** On March 1st, 1847 Lamé claimed, at a meeting of the Académie des Sciences in Paris, that he had proved Fermat's Last Theorem, that there are no non-zero integer solutions to

$$x^n + y^n = z^n$$

with $n \geq 3$. We can assume that x, y, z are pairwise coprime (else we can divide through by any common factor) and that n is an odd prime (since Fermat proved the case $n = 4$, and as an rsth power is also an rth power). Moreover, since n is odd we may permute x, y and $-z$ to guarantee that n does not divide z.

Now note that if $a_1, a_2, \ldots, a_k$ are pairwise coprime integers whose product is the nth power of an integer then, using the unique factorization theorem we can deduce that each a_j is the nth power of an integer. Lamé's idea was to reproduce the same argument for the Fermat equation. First he factored $z^n = x^n + y^n$ as $(x+y)(x+\zeta y)(x+\zeta^2 y)\ldots(x+\zeta^{n-1}y)$ where $\zeta = e^{2i\pi/n}$ is a primitive nth root of unity, and then proved that $(x+\zeta^i y, x+\zeta^j y) = 1$ whenever $i \neq j$. He therefore deduced that each $x+\zeta^j y$ is an nth power, and from such a wealth of surprising information deduced a contradiction.

Liouville spoke immediately after Lamé at the meeting and noted that there seemed to be a gap in the above proof. The assertion that whenever one has a product of pairwise coprime integral polynomials in ζ equalling an nth power, each of the polynomials is itself an nth power, requires proof. The analogous statement for integers relied on the unique factorization of integers, and it seemed to Liouville necessary to prove an analogous result in this case. Liouville also noted that even if unique factorization does hold, then all one can deduce is that each factor is a *unit* times an nth power where, by a unit, we mean a number that divides 1. There are just two units in the integers, namely 1 and -1, and these are both nth powers since n is odd. However there are other units in our setting; for example ζ^k for each k, $1 \leq k \leq n-1$, and more complicated examples like $\zeta + \overline{\zeta}$ since $(\zeta + \overline{\zeta})(\zeta + \zeta^5 + \zeta^9 + \cdots + \zeta^{2n-1}) = 1$, and these can often be shown to not be nth powers.

In fact the unique factorization assumption is false; Cauchy showed a couple of months later in 1847 that it fails for $n = 23$. Similar discussions had taken place at the Berlin Academy a year or two earlier, involving Dirichlet and Kummer, though the precise details of who thought what when, are not preserved. What we do know is that these discussions led Kummer to the development of an appropriate alternative theory of ideals, as we shall see in the next section, and he was able to use that to resurrect Lamé's proof of Fermat's last theorem for certain prime exponents n, the *regular primes*, as we will discuss a little later.

3. A general theory

3.1. **Ideals.** Again start with two integers $n \geq m > 0$ and let $\ell = n - m$. If r and s are any two integers then $mr + ns = \ell s + m(r+s)$, and if t and u are any two integers then $\ell t + mu = m(u-t) + nt$, so that the set of integral linear combinations of m and n is the same as the set of integral linear combinations of ℓ and m. Using this observation at each step in the Euclidean algorithm we discover that the set of integral linear combinations of m and n is the same as the set of integral multiples of the greatest common divisor of m and n.

This development led Kummer to a rich generalization of the notion of greatest common divisor. An integer can be identified by its set of multiples, and thus the greatest common multiple of m and n can be identified with the set of integral linear combinations of m and n. This is what can be generalized to other situations: instead of searching for the largest integer that divides every integer in a given set, we work with the set of integral linear combinations of our given set. Thus if A is our *ring of integers* (for examples, $\mathbb{Z}$, $\mathbb{Z}[t]$ and $\mathbb{Z}[\sqrt{2}] := \{a + b\sqrt{2} : \ a, b \in \mathbb{Z}\}$) then for any given $m_1, \ldots, m_r \in A$ we define the *ideal*

$$(m_1, \ldots, m_r) := \{a_1 m_1 + \cdots + a_r m_r : \ a_1, \ldots, a_r \in A\}.$$

The simplest example of an ideal, a *principal ideal*, is one that can be generated by just one element, in other words it is of the form (m) for some $m \in A$. We saw above that in $\mathbb{Z}$ any ideal generated by two integers can be written as an ideal generated by one integer, and is thus a principal ideal. Now any ideal in $\mathbb{Z}$ is of the form $(m_1, \dots, m_r)$ so, by induction on the number of generators, we can deduce that any ideal in $\mathbb{Z}$ is principal, and thus $\mathbb{Z}$ is a *principal ideal domain*.

Ideals in $\mathbb{Z}[\sqrt{d}]$ are not always principal, for example the ideal $(2, \sqrt{-6})$.[9] However, every ideal in $\mathbb{Z}[\sqrt{d}]$ can be written in terms of at most two generators and, in fact, all elements of the ideal are integral linear combinations of those two generators: Let $r+s\sqrt{d}$ be an element of our ideal with $s > 0$ minimal. We claim that s divides n for every other element $m+n\sqrt{d}$ of the ideal, for if not then write $n = qs+r$ where $1 \leq r \leq s-1$ and so $(m+n\sqrt{d}) - q(a+s\sqrt{d}) = (m-aq) + r\sqrt{d}$ is in our ideal, contradicting the minimality of s. Hence every other element of the ideal is an integral multiple of $r+s\sqrt{d}$ plus some integer, that is an integral linear combination of $r+s\sqrt{d}$ and the greatest common divisor of those integers, call it m. Now $sd + r\sqrt{d} = \sqrt{d}(r+s\sqrt{d})$ is also in the ideal, as is $m\sqrt{d}$, so that s divides both r and m. Writing $m = as$ and $r = bs$ we find that the elements of the ideal are precisely s times the integer linear combinations of a and $b+\sqrt{d}$.

An ideal containing a unit must be the whole ring. We multiply two ideals I and J by taking $IJ = \{ij : \; i \in I, \; j \in J\}$; a set of generators for IJ can be obtained by multiplying together the generators of I and of J. Note that IJ is a subset of both I and J (as an example in $\mathbb{Z}$, note that the set of integer multiples of 15 is a subset of the integer multiples of 3, and of 5). A *prime* ideal is an ideal which cannot be factored into two strictly larger ideals.[10]

Kummer's remarkable result is that, even though there is not a unique factorization theorem for the ring of integers of every field,[11] there is in fact a unique factorization theorem for the set of ideals of the ring of integers of every field. In other words every ideal may be written in a unique way as a product of prime ideals. This notion is essential to be able to work with the arithmetic of number fields. In our example above note that

$$(2, \sqrt{-6})^2 = (2 \cdot 2, 2 \cdot \sqrt{-6}, \sqrt{-6} \cdot \sqrt{-6}) = (4, 2\sqrt{-6}, 6) = (2, 2\sqrt{-6}) = (2),$$

[9] For if this equals $(a + b\sqrt{-6})$ then $a^2 + 6b^2$ divides 2, which implies $a = \pm 1, \; b = 0$, which is impossible since 1 is not a linear combination of 2 and $\sqrt{-6}$.

[10] In $\mathbb{Z}$ we artificially ignore factorizations like $5 = 5 \times 1$; working with ideals this corresponds to $(5) = (5) \times (1) = (5) \times \mathbb{Z}$ but here only one of the two ideals is strictly larger than (5).

[11] We will define "the ring of integers" of a number field in the next section.

and similarly $(3, \sqrt{-6})^2 = (3)$, and hence we have factorization of the ideal (6) in $\mathbb{Z}[\sqrt{-6}]$ into prime ideals:

$$(6) = (2) \cdot (3) = (2, \sqrt{-6})^2\ (3, \sqrt{-6})^2.$$

On the other hand the ring $\mathbb{Z}[\sqrt{6}]$ has unique factorization and therefore the ideal (6) factors into prime ideals as

$$(6) = (2+\sqrt{6})(2-\sqrt{6})(3+\sqrt{6})(3-\sqrt{6});$$

but note that we cannot deduce that the product of the numbers $(2+\sqrt{6})(2-\sqrt{6})(3+\sqrt{6})(3-\sqrt{6})$ equals 6; in fact it equals -6. What explains this difference of a minus sign? In general if we have two principal ideals $(\alpha) = (\beta)$ then $\beta \in (\alpha)$ and so β is a multiple of α, and vice-versa. So we have $\alpha = u\beta$ where both u and $1/u$ are in our ring. If the ring is $\mathbb{Z}$ then the only possibilities for u are 1 and -1, and this is the same for $\mathbb{Z}[\sqrt{-6}]$. However there can be many more possibilities in a more complicated number field: for example in $\mathbb{Z}[\sqrt{6}]$ we can have $u = 5+2\sqrt{6}$ since $1/u = 5-2\sqrt{6}$, and indeed if $u = \pm(5+2\sqrt{6})^k$ for some integer k then $1/u = \pm(5-2\sqrt{6})^k$. Such numbers are units and we need to better understand them.

3.2. **Number fields, algebraic integers and units.** We have used the term "ring of integers" without definition, something we now need to correct. A fraction can be thought of as a root of a linear equation with integer coefficients; what distinguishes integers is that the linear equation is monic. This viewpoint generalizes nicely: An *algebraic number* α is the root of an irreducible polynomial with integer coefficients (which is called the *minimum polynomial* for α), and an *algebraic integer* is an algebraic number for which the minimum polynomial is monic. It is worth noting that if α is an algebraic number then there exists a positive integer m such that $m\alpha$ is an algebraic integer. Also that the sum and the product of two algebraic integers is also an algebraic integer.

For a given finite set $\{\alpha_1, \alpha_2, \dots, \alpha_k\}$ of algebraic numbers, the set of rational functions, with integer coefficients, involving $\alpha_1, \alpha_2, \dots, \alpha_k$ is called a *number field*, denoted $\mathbb{Q}(\alpha_1, \alpha_2, \dots, \alpha_k)$.[12] Thus $\mathbb{Q}(\sqrt{d})$, the set of rational functions in $\sqrt{d}$, is a number field, called a *quadratic field*. We can assume that d is squarefree since $\sqrt{b^2d} = b\sqrt{d}$. The integers of this field are the algebraic integers in the field. Note that by multiplying top and bottom of $(r+s\sqrt{d})/((u+v\sqrt{d})$ by $u - v\sqrt{d}$ we may assume that all of the elements of $\mathbb{Q}(\sqrt{d})$ take the form $(r+s\sqrt{d})/t$ for integers r, s, t with $(r, s, t) = 1$ and $t > 0$. This is a root of $t^2x^2 - 2rtx + r^2 - ds^2$; and is thus an algebraic integer if and only

[12] A *rational function* is the quotient of two polynomials.

if t^2 divides $(2rt, r^2-ds^2)$. In this case no odd prime p can divide t, or else p divides r and p divides s as d is squarefree; and similarly 4 cannot divide t. Therefore $t=1$, or $t=2$ with r and s odd and $d \equiv 1 \pmod 4$, and so the ring of integers of $\mathbb{Q}(\sqrt{d})$ is $\mathbb{Z}[\sqrt{d}]$, the set of integer linear combinations of 1 and $\sqrt{d}$, if $d \equiv 2$ or $3 \pmod 4$, and $\mathbb{Z}[\frac{1+\sqrt{d}}{2}]$, the set of integer linear combinations of 1 and $\frac{1+\sqrt{d}}{2}$ if $d \equiv 1 \pmod 4$.

To determine the units we must find those algebraic integers u such that $1/u$ is also an algebraic integer, in other words the units are the roots of irreducible monic polynomials with constant term 1 or -1. Therefore in $\mathbb{Q}(\sqrt{d})$ we are looking for $r+s\sqrt{d}$ with r, s integers such that $r^2-ds^2=1$ or -1, and, if $d \equiv 1 \pmod 4$ for those $\frac{r+s\sqrt{d}}{2}$ with $r-s$ even such that $r^2-ds^2=4$ or -4. For examples $5+2\sqrt{6},\ 1+\sqrt{2},\ \frac{1+\sqrt{-3}}{2}$ and $\frac{1+\sqrt{5}}{2}$. We deduce that there cannot be a unit in $\mathbb{Q}(\sqrt{d})$ other than 1 or -1 when $d<0$, except for when $d=-3$ and $d=-1$. We will see later that there is always a unit other than 1 or -1 when d is positive and squarefree.

If u and u' are units then uu' and u/u' are also units, so that the units in a given number field form a multiplicative group. The units of finite order are roots of unity, the rest have infinite order. The unit group is therefore of the form $\mathcal{T} \oplus \mathbb{Z}^r$ where $\mathcal{T}$, the *torsion subgroup* of elements of finite order, is a finite cyclic group, and r is the *unit rank*, which describes the set of units of infinite order in the field. The units of finite order in quadratic fields are 1 and -1, as well as $\pm i \in \mathbb{Q}(\sqrt{-1})$, and $\frac{\pm 1 \pm \sqrt{-3}}{2} \in \mathbb{Q}(\sqrt{-3})$. Imaginary quadratic fields have unit rank zero, and real quadratic fields have unit rank one; thus for example the elements of the unit group in $\mathbb{Q}(\sqrt{6})$ are $\pm(5+2\sqrt{6})^k,\ k \in \mathbb{Z}$, which has the structure $\mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}$.

3.3. **The Gaussian integers.** The *Gaussian integers* are the set of algebraic integers in $\mathbb{Q}(i)$ where $i=\sqrt{-1}$, which turns out to be $\mathbb{Z}[i]$, the integer linear combinations of 1 and i. This is a ring with unique factorization (up to units), and one might ask how each prime in $\mathbb{Z}$ factors here? The first thing to note is that if $p=(a+ib)(a-ib)$ then $p=a^2+b^2$. Since squares can only be 0 or 1 mod 4, it is evident that $p \not\equiv 3 \pmod 4$. We have that $2=1+1=(1+i)(1-i)$ so that 2 factors, hence the question remains only for the primes $\equiv 1 \pmod 4$. Fermat showed that every such prime is the sum of two squares; we will do so assuming the easily proved fact that -1 is a square modulo prime p whenever $p \equiv 1 \pmod 4$:[13] Suppose that t is an integer for which $t^2+1 \equiv 0 \pmod p$. The set $\{i+jt:\ 0 \le i,j \le [\sqrt{p}]\}$ has $([\sqrt{p}]+1)^2 > p$ elements, and so two must be congruent mod p, say $i+jt \equiv I+Jt \pmod p$. Taking $a=i-I$ and $b=j-J$ we have that a and b are not both 0, and $|a|,|b| < \sqrt{p}$, so that $0 < a^2+b^2 < 2p$.

[13] Let $x=(\frac{p-1}{2})!$ so that $(p-1)(p-2)\dots(p-\frac{p-1}{2}) \equiv (-1)^{(p-1)/2}x \equiv x \pmod p$, and therefore $x^2 \equiv (p-1)! \equiv -1 \pmod p$ by Wilson's theorem.

Moreover $a \equiv -bt \pmod p$ so that $a^2 \equiv b^2t^2 \equiv -b^2 \pmod p$ and thus p divides a^2+b^2. These two facts imply that $a^2+b^2=p$.

One thing is left to investigate: if p does factor into two parts in $\mathbb{Z}[i]$, are these parts distinct? In other words, if $p=(a+ib)(a-ib)$ is it possible that $a+ib=u(a-ib)$ for some unit u? The only units of $\mathbb{Z}[i]$ are $1,-1,i,-i$ leading to $b=0, a=0, a=b, a=-b$ respectively. We deduce that $2=i(1-i)^2$ is the only prime that has repeated factors.

So, to summarize, we have proved that prime p factors into two primes in $\mathbb{Z}[i]$ if and only if $p \equiv 1 \pmod 4$, in which case the prime factors are distinct, or $p=2$ in which case the ideal (2) is the square $(1-i)^2$.

This all generalizes rather nicely to $\mathbb{Q}(\sqrt{d})$. The ideal (p) for odd rational prime p factors into two prime ideals in $\mathbb{Q}(\sqrt{d})$ if and only if d is a square mod p. In fact, d is a square mod p if and only if p belongs to certain arithmetic progressions mod $4d$. If p does not divide $4d$ then the two prime ideals in the factorization of (p) are distinct. In this case if p divides d then the ideal (p) is the square of a prime ideal.[14]

3.4. **Factoring a prime p in a given number field.** So how do we factor p in a ring of integers, say $\mathbb{Z}(\alpha)$ (which is the set of polynomials, with integer coefficients, in algebraic integer α)? Kronecker made the surprising observation that this is tantamount to factoring $f(x) \pmod p$, where $f(x)$ is the minimum polynomial of α (and remember that minimum polynomials are irreducible): Suppose that the (unique) factorization of $f(x) \pmod p$ is

$$f(x) \equiv g_1(x)^{e_1} g_2(x)^{e_2} \dots g_k(x)^{e_k} \pmod p$$

where the $g_j(x)$ are distinct irreducible polynomials mod p, and the e_j are positive integers. Then p divides $g_1(\alpha)^{e_1} g_2(\alpha)^{e_2} \dots g_k(\alpha)^{e_k}$, and $(g_i(\alpha), g_j(\alpha), p)=1$ for $i \neq j$, and so

$$(p) = (p, g_1(\alpha)^{e_1} g_2(\alpha)^{e_2} \dots g_k(\alpha)^{e_k}) = (p, g_1(\alpha)^{e_1})(p, g_2(\alpha)^{e_2}) \dots (p, g_k(\alpha)^{e_k}).$$

If p does not divide the *discriminant*[15] of f then all the e_js equal 1 and so

$$(p) = (p, g_1(\alpha))(p, g_2(\alpha)) \dots (p, g_k(\alpha)),$$

the desired factorization into prime ideals. A similar, but more complicated, result holds when p divides the discriminant of f.

[14] Otherwise, how the prime 2 factors, requires an unenlightening case-by-case analysis.

[15] The discriminant of a polynomial $f(x)$ is more-or-less the greatest common divisor of $f(x)$ and $f'(x)$ in the ring $\mathbb{Z}[x]$ (which is defined to be the minimum possible outcome of the Euclidean algorithm in this setting). Note that this value must be divisible by any prime p for which $f(x) \pmod p$ has repeated roots.

One beautiful example comes in taking the pth cyclotomic field, $\mathbb{Q}(\zeta_p)$, where $\zeta = \zeta_p = e^{2i\pi/p}$ is a primitive pth root of unity. This has minimal equation $(x^p-1)/(x-1)$, which is $\equiv (x-1)^{p-1} \pmod p$ since $(x-1)^p \equiv x^p-1 \pmod p$. Thus $(p) = (p, (1-\zeta)^{p-1})$ and one can deduce that $(p) = (1-\zeta)^{p-1}$; that is (p) factors into principal ideals, and so the two sides differ multiplicatively by a unit. Finding a nice presentation of that unit, for example its minimal polynomial, is not an easy task. The same proof yields that $(p) = (1-\zeta^k)^{p-1}$, for any integer k, $1 \leq k \leq p-1$ and so $(1-\zeta^k)/(1-\zeta)$ is a unit.

In Lamé's putative proof of Fermat's Last Theorem, discussed above, he determined $(x+\zeta^i y, x+\zeta^j y)$ where $(x,y)=1$: This ideal contains the elements $(x+\zeta^i y)-(x+\zeta^j y) = (\zeta^i - \zeta^j)y$ and $\zeta^j(x+\zeta^i y) - \zeta^i(x+\zeta^j y) = (\zeta^j - \zeta^i)x$, and therefore $(\zeta^i-\zeta^j)(x,y) = \zeta^j(\zeta^{i-j}-1)$. We just saw that $(1-\zeta^{i-j})/(1-\zeta)$ is a unit, and so our ideal contains the element $(1-\zeta)$ as well as $(x+\zeta^i y) + (1-\zeta)y(1-\zeta^i)/(1-\zeta) = x+y$, and so we deduce that $(x+\zeta^i y, x+\zeta^j y) = (1-\zeta, x+y)$. As $1-\zeta$ divides p thus our ideal divides $(p, x+y)$ which equals 1 if p does not divide z.

4. Groups

4.1. **Constructing units.** Suppose that d is a squarefree integer > 1. If $d \equiv 2$ or $3 \pmod 4$ then $\mathbb{Z}[\sqrt{d}]$ is the ring of integers of $\mathbb{Q}(\sqrt{d})$, so if $u = a + b\sqrt{d}$ is a unit then $a^2 - db^2 = 1$ or -1, and hence $(2a)^2 - d(2b)^2 = 4$ or -4. If $d \equiv 1 \pmod 4$ then $\mathbb{Z}[\frac{1+\sqrt{d}}{2}]$ is the ring of integers of $\mathbb{Q}(\sqrt{d})$, so if $u = (a+b\sqrt{d})/2$, with $a-b$ even, is a unit then $a^2 - db^2 = 4$ or -4. Either way we are searching for solutions to the *Pell equation*

$$x^2 - dy^2 = \pm 4,$$

where x and y are integers with $x-y$ even. We are not interested in the solutions with $y=0$ (which correspond to the units ± 1). Let (u,v) be the solution with $\epsilon_d = \frac{u+v\sqrt{d}}{2}$ smallest but > 1; we claim that every solution with $\frac{x+y\sqrt{d}}{2} > 1$ takes the form

$$\left(\frac{x+y\sqrt{d}}{2}\right) = \left(\frac{u+v\sqrt{d}}{2}\right)^k$$

for some integer $k \geq 1$. If not, let $\frac{x+y\sqrt{d}}{2}$ be the smallest counterexample. We must have $\frac{x+y\sqrt{d}}{2} > \frac{u+v\sqrt{d}}{2}$ by definition of u, v, but then $\pm\frac{x+y\sqrt{d}}{2} \cdot \frac{u-v\sqrt{d}}{2}$ where the '±' is chosen to have the same sign as $u^2 - dv^2$, is a smaller counterexample, giving a contradiction. The solution u, v is known as the *fundamental solution* to Pell's equation and every unit of $\mathbb{Q}(\sqrt{d})$ can be uniquely written in the form $\pm\epsilon_d^k$ for some integer k.

A real number α has a continued fraction of finite length if and only if α is rational. A real number is in a quadratic field and thus of the form $(b+\sqrt{d})/2a$ where a, b and d are integers if and only if its continued fraction is eventually periodic (see, e.g.

[1]); that is, there exists an integer m such that $a_{n+m} = a_n$ for all sufficiently large n. When $\alpha = [a_0, a_1, \dots]$ is *purely periodic*, that is $a_{n+m} = a_n$ for all $n \geq 0$, then $\alpha = [a_0, a_1, \dots, a_{m-1}, \alpha]$ and so for some $\lambda \neq 0$ we have

$$\lambda \begin{pmatrix} \alpha \\ 1 \end{pmatrix} = \begin{pmatrix} a_0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a_1 & 1 \\ 1 & 0 \end{pmatrix} \cdots \begin{pmatrix} a_{m-1} & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ 1 \end{pmatrix} = \begin{pmatrix} p_{m-1} & p_{m-2} \\ q_{m-1} & q_{m-2} \end{pmatrix} \begin{pmatrix} \alpha \\ 1 \end{pmatrix},$$

from which we deduce that $q_{m-1}\alpha^2 + (q_{m-2} - p_{m-1})\alpha - p_{m-2} = 0$. The continued fraction for $\sqrt{d} + [\sqrt{d}]$ is purely periodic (see the end of section 6.4 of [1]), and thus $a_{n+m} = a_n$ for all $n \geq 1$ in the continued fraction for $\sqrt{d}$. Therefore if $\alpha_r = [a_r, a_{r+1}, \dots]$ we find that $\alpha_{m+1} = \alpha_1 = 1/(\sqrt{d} - [\sqrt{d}])$ and, proceeding as above,

$$\begin{pmatrix} \sqrt{d} \\ 1 \end{pmatrix} = \lambda \begin{pmatrix} p_m & p_{m-1} \\ q_m & q_{m-1} \end{pmatrix} \begin{pmatrix} \alpha_{m+1} \\ 1 \end{pmatrix} = \lambda' \begin{pmatrix} p_m & p_{m-1} \\ q_m & q_{m-1} \end{pmatrix} \begin{pmatrix} 1 \\ \sqrt{d} - [\sqrt{d}] \end{pmatrix}.$$

Expanding out and comparing the (rational integer) coefficients of 1 and $\sqrt{d}$ in the resulting expression, we deduce that

$$p_{m-1} = q_m - [\sqrt{d}]q_{m-1}, \text{ and } dq_{m-1} = p_m - [\sqrt{d}]p_{m-1},$$

so that

$$p_{m-1}^2 - dq_{m-1}^2 = p_{m-1}(q_m - [\sqrt{d}]q_{m-1}) - q_{m-1}(p_m - [\sqrt{d}]p_{m-1}) = (-1)^m,$$

yielding a solution to Pell's equation. This technique can be dated back to Brahmagupta, an Indian mathematician who lived at the end of the 6th century, and was perhaps known even earlier.

Archimedes' cattle problem, is a 22 line epigram, which he sent to the mathematicians of Alexandria, in 251 B.C. It begins by asking the reader to find the numbers of white, black, blue and spotted bulls and cows, where these numbers satisfy eight given linear equations. Archimedes writes "if you can solve the problem up to this point no one will call you ignorant, but this does not yet make you an expert". He then gives two further equations: In one a certain sum of the variables equals a square; in the other a different sum of the variables equals a triangular number. Archimedes then adds: "if you can solve this now, then you win the prize for supreme wisdom". It can be shown by the theory of Pell's equation that Archimedes' cattle problem is equivalent to finding the 2329th smallest solution to $x^2 - dy^2 = 1$ with $d = 4729494$ and y divisible by 9314, so that the total number of cattle is an integer with 206545 decimal digits. Presumably Archimedes' understood the difficulty of his problem because he had a firm grasp of the mathematics behind Pell's equation (see [10] for more on this charming question).

4.2. **Irreducibles.** The ideal (5) factors in the field $\mathbb{Q}(\sqrt{19})$ as $(3-\sqrt{19},5)(3+\sqrt{19},5)$, that is into two non-principal ideals. If we restrict our attention just to the algebraic integers of the field then 5 cannot be factored, that is 5 is *irreducible* but not prime in $\mathbb{Q}(\sqrt{19})$.

One might ask whether there are irreducibles in a given number field that can be split into arbitrarily many prime factors, or whether there is a bound on the possible number? If there is a bound then this is some kind of measure on how far the given number field is away from having unique factorization. In fact there is a bound, and understanding this leads us into our next topic, the class group:

4.3. **The class group.** We wish to measure how far away ideals are from being principal in a given field K. To do this the modern algebraist studies "ideals modulo principal ideals"; by this we mean that two ideals are considered to be the same in this setting if they differ, multiplicatively, by a principal ideal. More precisely we say that ideals I and J are *equivalent* if there exist algebraic integers α and β, of K, for which $(\alpha)I = (\beta)J$. Thus any two principal ideals are equivalent. Any set of ideals that are equivalent to one another is an *ideal class*; so the principal ideals form the *principal ideal class*.

For example in the field $\mathbb{Q}(\sqrt{-5})$ we have

$$(1-\sqrt{-5}) \times (2, 1+\sqrt{-5}) = (2(1-\sqrt{-5}), 6) = (2) \times ((1-\sqrt{-5}), 3)$$

so that the ideals $(2, 1+\sqrt{-5})$ and $((1-\sqrt{-5}), 3)$ are equivalent.

Notice that if ideals I and J are equivalent to ideals A and B, respectively, then IJ is equivalent to AB. Thus we can define multiplication of ideal classes via multiplication of ideals, and this multiplication is commutative. Evidently the principal ideal class is the identity in this multiplication. If $K = \mathbb{Q}(\sqrt{-d})$ then the product of any ideal with its complex conjugate[16] gives a principal ideal, so that every ideal class has an inverse, and hence the ideals form an abelian group, called the *ideal class group*. If $K = \mathbb{Q}(\sqrt{d})$ then we obtain the conjugate ideal via the map $\sqrt{d} \to -\sqrt{d}$, and an analogous, though more involved, construction of an inverse ideal class works for ideal classes in any ring of integers.

How many distinct ideal classes are there in the ring of integers of a given number field? (The *class number* is defined to be the number of distinct ideal classes.) That is, how large is the class group? If there is just one ideal class, that is the class number $h(K) = 1$, then all of the ideals are principal, and thus we have a *principal ideal domain* which implies that we have unique factorization. If $h(K) \neq 1$ then factorization is not unique. The first question to address is whether $h(K)$ is bounded, or whether it can

[16] The *complex conjugate* of an ideal I is the ideal $\overline{I} := \{\overline{z} : \ z \in I\}$.

ever be infinite? If $d > 0$ is squarefree then we can use Gauss's algorithm, modelled on Euclid's algorithm, to prove that the class number of $\mathbb{Z}[\sqrt{-d}]$ is finite:[17]

In Euclid's algorithm we have two possible actions given integers n and m:

(i) Unless $n < m$ replace n by the least non-negative residue, n' of $n \pmod m$, that is, the residue in $[0, m)$. Evidently $(n, m) = (n', m)$. In the continued fraction algorithm for n/m this amounts to subtracting $[n/m]$ from n/m to obtain a number in $[0, 1)$.

(ii) If $n < m$ then we simply swap the two numbers, comparing m and n. Evidently $(n, m) = (m, n)$. In the continued fraction algorithm for n/m this amounts to inverting n/m, replacing it by m/n.

In Gauss's algorithm we begin with two generators a and $b + \sqrt{-d}$ of an ideal in $\mathbb{Q}(\sqrt{-d})$, for squarefree $-d < 0$: note again that a divides $b^2 + d$. Here are Gauss's two analogous actions:

(i) Unless $-a/2 < b \leq a/2$ replace b by the least residue, in absolute value, of $b \pmod a$, that is the residue b' in $(-a/2, a/2]$. Evidently $(a, b + \sqrt{-d}) = (a, b' + \sqrt{-d})$

(ii) If $-a/2 < b \leq a/2$ then we invert $(b + \sqrt{-d})/a$, writing $b^2 + d = ac$ for some integer c, to obtain $a/(b+\sqrt{-d}) = (b-\sqrt{-d})/c$. Evidently $(b-\sqrt{-d}) \times (a, b+\sqrt{-d}) = (a(b-\sqrt{-d}), b^2+d) = (a) \times (b-\sqrt{-d}, c)$ so that the ideals $(a, b+\sqrt{-d})$ and $(c, b-\sqrt{-d})$ are equivalent.

Note that if $a > \sqrt{4d/3}$ then $ca = b^2 + d < a^2$, that is $c < a$; in other words Gauss's algorithm, like Euclid's algorithm, reduces the size of the numbers involved, at least if the numbers are large enough. Moreover this shows that each equivalence class of ideals contains an ideal $(a, b + \sqrt{-d})$ with $|2b| \leq a \leq \sqrt{4d/3}$, and so there are only finitely many possibilities; that is the class number is indeed finite.

The *norm* of the ideal $(a, b + \sqrt{-d})$ is $|a|$; Gauss's proof shows that every ideal class contains an ideal of norm $\leq \sqrt{4d/3}$. This proof generalizes to establish that in any number field every equivalence class contains some ideal with norm beneath a certain bound that depends on the field, and therefore the class group is finite.

How big is $h(K)$ typically? Much depends on what type of field K is. For K of the form $\mathbb{Q}(\sqrt{d})$ we have that $h(K)$ is typically around $\sqrt{|d|}$ when d is negative, but is typically bounded when d is positive. Gauss asked an important question in each case:

- Is it true that there are infinitely many squarefree $d > 0$ for which the class number is one?
- Are there negative squarefree d for which the class number is one, other than the nine values given in the list $-1, -2, -3, -7, -11, -19, -43, -67, -163$?

[17] This is the ring of integers of $\mathbb{Q}(\sqrt{-d})$ if $-d \equiv 2$ or $3 \pmod 4$, and a subring if $-d \equiv 1 \pmod 4$. This algorithm can easily be modified for ideals in the full ring of integers, in the latter case.

The first question remains completely open. The quest to resolve the second question set the tone for twentieth century number theory perhaps more than any other problem. In the 1930s it was shown that there are no more than ten elements on the list, though the proof, by its very nature, cannot be modified to determine whether there is indeed a missing tenth d. In the 1950s, Heegner showed that there is no tenth field by a proof that was not fully believed at the time; though nowadays we know that Heegner was correct and the technique he created to prove this result is now central to arithmetic geometry. In the 1960s Baker and Stark came up with quite different, and widely accepted proofs that there is no tenth field. In the 1980s Goldfeld, Gross and Zagier showed how one can find all squarefree $-d < 0$ with any given class number, be it 1, 2 or whatever.

Armed with the class group we now show that, for any number field K, an irreducible α in K can have no more than $B(K)$ prime factors, for some bound $B(K)$ depending only on K. We use Lagrange's result that if G is a finite group and $g \in G$ then $g^{\#G} = 1$, where 1 is the identity in G. If the factorization of the ideal (α) into prime ideals is $\mathcal{P}_1\mathcal{P}_2 \dots \mathcal{P}_k$ then we claim that there are no more than $h(K) - 1$ of $\mathcal{P}_j$ in any given ideal class, for if there are $h(K)$ then the product of these ideals is principal, say, (β) and we can write $\alpha = \beta\gamma$ for algebraic integers β, γ, so that α is reducible. Therefore $B(K) \leq (h(K)-1)^2$, and Davenport asked the still open question as to the best possible such bound $B(K)$ for each number field K. In fact $B(K) = B(G)$ where G is the class group, and $B(G)$ is the largest possible number of elements of an abelian group G such that their product is the identity but the product of the elements of any proper subset is never the identity.

4.4. **Equations as examples.** We will now find all integer solutions to the equation

$$x^2 + 2 = y^3.$$

Note first that y is odd, or else x is even and we have $0 + 2 \equiv 0 \pmod 4$ which is impossible. We have already seen that $\mathbb{Q}(\sqrt{-2})$ has class number one, and thus has unique factorization, and that its only units are 1 and -1 which are both cubes of themselves. Now $x^2 + 2 = (x + \sqrt{-2})(x - \sqrt{-2})$ and the two factors are coprime (since $(y, 2) = 1$); therefore $x + \sqrt{-2}$ and $x - \sqrt{-2}$ must both be cubes of elements of $\mathbb{Z}[\sqrt{-2}]$. Now if $x + \sqrt{-2} = (u + v\sqrt{-2})^3$ for some integers u and v then $3u^2v - 2v^3 = 1$ so that $v = \pm 1$ and therefore $3u^2 = 2 + v$. This yields $v = 1,\ u = \pm 1$ and therefore $x = \pm 5$ and $y = 3$.

Let us now apply the same procedure to find all integer solutions to the equation

$$x^2 + 19 = y^3.$$

First note that 19 cannot divide y (or else it divides x and the equation is impossible mod 19^2), and that 2 cannot divide y as there is no solution possible to $x^2+19 \equiv 0 \pmod 8$. Now $x^2+19 = (x+\sqrt{-19})(x-\sqrt{-19})$ and the two factors are coprime (since $(y,38)=1$) so, as ideals, must each equal the cube of an ideal. Now the class number for the ring of integers of $\mathbb{Q}(\sqrt{-19})$ is one, and therefore all ideals are principal. Moreover the only units are 1 and -1 and thus both cubes. Therefore if $x+\sqrt{-19} = (u+v\sqrt{-19})^3$ for some rational integers u and v then the coefficient of $\sqrt{-19}$ is $1 = 3u^2v - 19v^3$, so that $v = \pm 1$ and therefore $3u^2 = 19+v$ which is impossible. Therefore there are no solutions to our equation. **However** this is wrong since $18^2+19 = 7^3$.

So where does our purported proof go wrong? One version of what went wrong is that the ring of integers is not $\mathbb{Z}[\sqrt{-19}]$ but rather $\mathbb{Z}[\frac{1+\sqrt{-19}}{2}]$, whence we should have written $x+\sqrt{-19} = (\frac{u+v\sqrt{-19}}{2})^3$ for some integers u and v with $u-v$ even. Then $8 = 3u^2v - 19v^3$ to which we find the unique solution $u=3, v=1$ and recover the only solution $(18,7)$ to the equation displayed above.

If we had chosen to solve the equation via the arithmetic of the ring $\mathbb{Z}[\sqrt{-19}]$ instead, then we would have run into another problem; the class number of this ring is 3 which means that the cube root of a principal ideal may well not be principal, and thus new complications arise.

We will work with other Diophantine questions below.

5. Quadratic forms, Ideals, and Transformations

5.1. **Different perspectives on reduction.** Suppose that $d \equiv 1 \pmod 4$. By the method developed earlier one can prove that all ideals of $\mathbb{Z}[\frac{1+\sqrt{d}}{2}]$ can be written as $(a, \frac{b+\sqrt{d}}{2})$ where a is the norm of the ideal and $b^2-d = 4ac$ for some integer c, and all the elements are integral linear combinations of the two generators, that is $\{ax + (\frac{b+\sqrt{d}}{2})y : x,y \in \mathbb{Z}\}$. Now we can associate to any conjugate pair of such forms $ax + (\frac{b+\sqrt{d}}{2})y$ and $ax + (\frac{b-\sqrt{d}}{2})y$, their product divided by their norm a, that is $ax^2+bxy+cy^2$.

Thus there is a 2-to-1 map from the ideals of $\mathbb{Z}[\frac{1+\sqrt{d}}{2}]$ to the *binary quadratic forms* $f(x,y) = ax^2+bxy+cy^2$ of discriminant $b^2-4ac = d$. We shall be interested in understanding the set of integers n *represented* by f, that is for which there exists integers u and v such that $n = au^2+buv+cv^2$. Now let $b' = b+2ak$ be the least residue, in absolute value, of $b \pmod{2a}$, let $c' = f(1,k)$ and $g(x,y) = ax^2+b'xy+c'y^2$. Then $f(u,v) = g(u-kv, v)$ so that f and g represent the same integers, and thus f is *equivalent* to g. Transforming from f to g is analogous to the first step in Gauss's algorithm discussed above. The second step in Gauss's algorithm has a much better description in this context, simply by mapping f to h where $h(x,y) = cx^2-bxy+ay^2$,

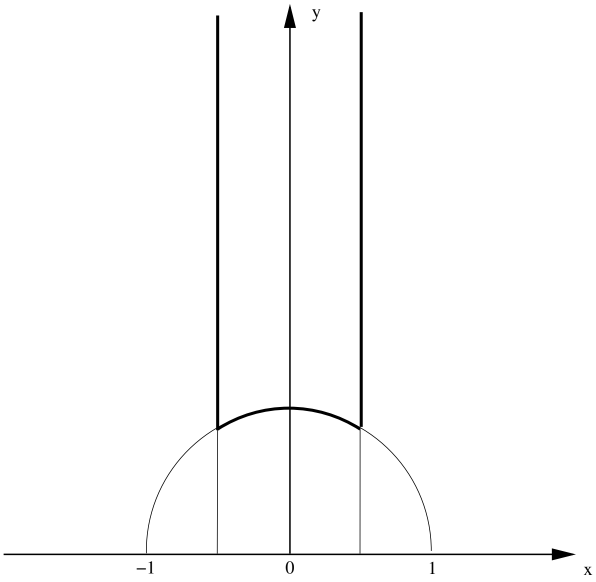

FIGURE 1. Fundamental domain of $\mathrm{SL}(2,\mathbb{Z})$ in the complex upper half plane.

since $f(u,v) = h(v,-u)$, and thus f is equivalent to h. This algorithm is what actually appeared in Gauss's [5]; the description given above, with ideals, first appeared in subsequent work of Dirichlet.

For a third equivalent description when $d < 0$, consider the complex number $z = \frac{b+\sqrt{d}}{2a}$ in the upper half of the complex plane. For the first part of Gauss's algorithm we map $z \to z' = z + k$ so that $-\frac{1}{2} < \mathrm{Im}z' \leq \frac{1}{2}$. For the second part of Gauss's algorithm, if $|z| < 1$ then we map $z \to z' = -1/z$ so that $|z'| > 1$. The algorithm terminates when z is in the *fundamental domain* $-\frac{1}{2} < \mathrm{Im}z \leq \frac{1}{2}$ with $|z| \geq 1$ (as in Figure 1). Note that the two steps of the algorithm are equivalent to applying the matrices

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \text{ and } \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \text{ to } \begin{pmatrix} z \\ 1 \end{pmatrix},$$

much like we saw when we discussed the Euclidean algorithm. In fact these two matrices generate multiplicatively $\mathrm{SL}(2,\mathbb{Z})$, the group of 2-by-2 matrices with integer entries of determinant one.

5.2. **Quadratic Forms.** In the previous section we saw how understanding what integers are represented by binary quadratic forms is tied in with unique factorization. It is a question of some interest to determine which integers are represented by a given quadratic form. For example Lagrange proved that every integer is the sum of four squares, and Ramanujan asked which quadratic forms represent all integers. Quite recently Bhargava and Hanke gave the easily applied criterion that a quadratic form with integer coefficients represents all the positive integers if and only if it represents each of the twenty-nine integers $1, 2, 3, 5, 6, 7, 10, 13, 14, 15, 17, 19, 21, 22, 23, 26, 29, 30, 31, 34, 35, 37, 42, 58, 93, 110, 145,$ 203 and 290.

6. Diophantine equations

6.1. **Fermat's Last Theorem, revisited.** In 1847 Kummer wrote to Liouville that he could prove Fermat's last theorem for prime exponent n assuming two properties of n. If we follow the discussion of Fermat's last theorem above, then we have that each ideal $(x + \zeta^j y)$ is the nth power of an ideal, call it B_j. This implies that if B_j is in the ideal class I then $I^n = 1$. Now consider only those primes n which do not divide the order of the class group of the ring of integers of $\mathbb{Q}(\zeta_n)$; we call these *regular primes*. Evidently if $I^n = 1$ then $I = 1$ and so B_j must be a principal ideal. This means that there exists an algebraic integer α_j such that $(x + \zeta^j y) = (\alpha_j)^n$; and therefore a unit u_j for which $x + \zeta^j y = u_j \alpha_j^n$. The second assumption that Kummer made implied that $u_j = v_j^n$ for some other unit v_j, and he was subsequently able to show that this second assumption always holds for regular primes n. Thus he deduced that each $x + \zeta^j y$ is the nth power of an algebraic integer, with which he was able to resurrect Lamé's argument and produce a contradiction. That is, Kummer proved Fermat's last theorem for regular prime exponents.

6.2. **Elliptic curves.** Suppose that f is a monic polynomial of degree 3 with integer coefficients, and no repeated roots. An *elliptic curve* E is the set of points on the curve $y^2 = f(x)$. We denote the points on E from the field K by $E(K)$. Poincaré proved that the points of $E(K)$ form an abelian group, three points summing to zero if they lie on a line, and asked whether the abelian group is finitely generated? That is, whether it is of the form $\mathcal{T} \oplus \mathbb{Z}^r$ where $\mathcal{T}$ is a finite torsion subgroup and r is some integer. This was proved by Mordell in the 1920s and we will now describe the part of his proof pertaining to unique factorization. We begin with the elliptic curve $y^2 = x(x-a)(x-b)$ where a and b are integers, and suppose that $(r/t^2, s/t^3)$ is a rational point on E, where $(r, t) = 1$. Therefore $r(r-at^2)(r-bt^2) = s^2$. Now $d_1 := (r, r-at^2) = (r, a)$, $d_2 := (r, r-bt^2) = (r, b)$ and $d_3 := (r - at^2, r - bt^2) = (r - at^2, b - a)$ are divisors of a, b and $b - a$, respectively, so that $r = d_1 d_2 u^2,\ r - at^2 = d_1 d_3 v^2,\ r - bt^2 = d_2 d_3 w^2$ for integers u, v, w.

More generally, suppose that f has no more than one rational root. Let K be the smallest field which contains all the roots of f. If we attempt to imitate the proof of the previous paragraph then we run into issues of unique factorization for the field K:

That is, the ideal generated by one of the factors linear in r and t^2 equals some ideal that is a divisor of the discriminant of f times the square of an ideal. Let us write this as $(\alpha) = DI^2$. Let I_0 be the ideal of smallest norm in the ideal class of I so that $I\overline{I_0}$ is principal, call it (β). Hence $(\alpha(NI_0)^2) = DI_0^2(\beta)^2$. Therefore DI_0^2 is a principal ideal, say (γ), coming from a finite set. Thus $\alpha = u\gamma\delta^2$ for some algebraic number δ and some unit u. If $u_1, \dots u_r$ is a basis for the units of K then for any unit u there is a subset S of $\{1, 2, \dots, r\}$ such that u is $\prod_{i\in S} u_i$ times the square of a unit; and so $\alpha = \gamma'\rho^2$ where γ' comes from some finite, computable set. We have proved a generalization of the result we had in the case that f splits into linear factors over $\mathbb{Q}$. This is essentially the argument that Mordell used in his proof of Poincaré's conjecture. Weil realized that Mordell did not work with the unit group U in the last step but rather with the finite quotient U/U^2, and that earlier he could have worked with C/C^2 rather than the class group C, and in fact one could have even started by working with $E(\mathbb{Q})/2E(\mathbb{Q})$. By doing so, Weil massively simplified Mordell's complicated proof, and ushered in the methods of modern arithmetic geometry.

7. Unique factorization, in practice

7.1. **Factoring.** The unique factorization theorem tells us that every integer can be factored into primes in a unique way, but it does not tell us how to do this in practice. As Gauss wrote in article 329 of [5]:

> "The problem of distinguishing prime numbers from composite numbers, and of resolving the latter into their prime factors is known to be one of the most important and useful in arithmetic. It has engaged the industry and wisdom of ancient and modern geometers to such an extent that it would be superfluous to discuss the problem at length. Nevertheless we must confess that all methods that have been proposed thus far are either restricted to very special cases or are so laborious and difficult that ... they try the patience of even the practiced calculator. And these methods do not apply at all to larger numbers ... *The dignity of the science itself seems to require that every possible means be explored for the solution of a problem so elegant and so celebrated* ... It is in the nature of the problem that *any* method will become more complicated as the numbers get larger. The techniques ... known ... require intolerable labor even for the most indefatigable calculator."

What Gauss wrote two hundred years ago is still true today. But now, more than for just the "dignity of the science itself", we study factoring because the seeming difficulty in factoring long integers is what keeps our electronic communications safe; that is, the impenetrability of the most commonly used cryptography is based on the fact that no one can factor 200 digit numbers quickly.

Since any composite number has a prime factor no larger than its square root, one can factor n by testing whether it is divisible by any number up to its square root. This is easily done for, say, $n = 1001$ or $n = 11041$, but how about for $n = 1234567890123456789$? This requires over a billion test divides, and if one were to try to factor a given 100 digit integer, which is the product of two 50 digit prime numbers, in this way then it would take longer than the remaining lifespan of our universe, even on an impossibly fast computer! One therefore needs a more sophisticated approach to handle large numbers.

Fermat came up with a method that is better suited to integers that are the product of two roughly equal primes, which he exhibited on the example $n = 2027651281$. First note that $r = 45029 = [\sqrt{n}]$ and that $n = r^2 + 40440$. Fermat's idea is to find j such that $(r+j)^2 - n$ is itself a square, say s^2, so that $n = (r+j+s)(r+j-s)$. He tried each j successively, and efficiently, as follows: $(r+1)^2 - n = (2r+1) - 40440 = 49619$ and this is not a square as it is $\equiv 19 \pmod{100}$. Next $(r+2)^2 - n = (2r+3) + 49619$ which he again ruled out mod 100. With each successive j he augmented the number by two more than the previous time, and ruled out the non-squares by modular arithmetic, eventually finding that $(r+12)^2 - n = 1020^2$, and deducing that $2027651281 = 44021 \times 46061$. Unfortunately Fermat's algorithm is very slow in the worst cases, as is a variant using binary quadratic forms, due to Gauss.

Modern factoring algorithms are mostly geared to working fast even in the worst case. Often they have the drawback that they are not guaranteed to always work, in that they may depend on a random number generator and the factorer might just be unlucky, but usually this can be organized in such a way that we would not expect to ever be so unlucky within the lifespan of the universe! The most efficient algorithm known is called the *number field sieve*,[18] which is a variation on the *quadratic sieve*, itself a variant of Fermat's original algorithm.

If n is composite and y is coprime to n then there are at least four solutions $x \pmod{n}$ to $x^2 \equiv y^2 \pmod{n}$, and so for at least half of these solutions we have that $(x-y,n)$ $(x+y,n)$ provides a factorization of n. In the different factoring algorithms we try to find such integers x and y (with $x^2 \equiv y^2 \pmod{n}$) by various methods. Typically

[18] I should perhaps say "publicly known". Most wealthy countries and corporations employ mathematicians secretly working on these problems, because of the cryptographic implications, and it may be that significant advances have been made behind closed doors!

one selects $a_1, a_2, \dots \pmod n$ and then takes b_j to be the least positive residue of $a_j^2 \pmod n$. One hopes to find a sequence of values $j_1 < j_2 < \cdots < j_k$ such that $b_{j_1} b_{j_2} \dots b_{j_k}$ is a square, say y^2, so that we have our solution above with $x = a_{j_1} a_{j_2} \dots a_{j_k}$. Once one has a process for generating the a_i, the key issue is to determine a subsequence of the b_j whose product is a square. We do this by working only with the b_j that have no prime factors $> B$, for some well-chosen B, and storing the factorizations of such b_j. Indeed if $b_j = \prod_{i=1}^{\ell} p_i^{c_{j,i}}$ then $b_{j_1} b_{j_2} \dots b_{j_k}$ is a square if and only if $\sum_{h=1}^{k} c_{j_h,i}$ is even for $i = 1, 2, \dots, \ell$. In other words if we create the matrix where the row corresponding to b_j is the vector of exponents $(c_{j,1}, \dots, c_{j,\ell})$, each taken mod 2, then we require a non-trivial subset of such rows whose sum is zero mod 2. This can be found efficiently using Gaussian elimination mod 2. Next we have to consider how to select the a_j. One way is to pick random integers, another, the values of polynomials. Early researchers found that numbers related to the continued fraction of $\sqrt{n}$ worked well. Each of these algorithms work in roughly $e^{\sqrt{d}}$ steps, where d is the number of decimal digits of n, a marked improvement on earlier algorithms that took more like e^d steps.

In the number field sieve one tries to imitate the quadratic sieve in number fields that are cleverly chosen to make the algorithm much more efficient. The above argument can be translated into this setting giving a running time of around $e^{d^{1/3}}$ steps. Most interesting for us is the step in which we factor b_j into small primes: First we factor the ideal (b) into prime ideals of small norm, and then we create a factorization of the algebraic integer b, proceeding much as we did in the previous section, taking account of the class group and unit group of the field. Moreover, just as in Weil's work, we can restrict our attention to C/C^2 and U/U^2, an observation that makes this algorithm practical.

7.2. **Cryptography.** Cryptographic protocols used to be based on complicated combinatorial schemes, and the safety of the secret message rested on keeping the key secret, because whoever had the key could easily invert it. In the mid-70s interest grew in finding *one-way functions* in which knowing the function did not help, in practice, in finding its inverse. Thus, a cryptographic protocol based on such a function would mean that one's enemy knowing the key would not, in practice, help them determine how to decode an encoded message. A candidate for such a one way function is multiplication: It is easy to multiply together two large primes, but not so easy to recover the two large primes from their product, as we discussed in the previous section. Rivest, Shamir and Adleman came up with a simple cryptographic protocol which we believe can be broken when properly implemented if and only if one can factor large numbers rapidly.

Is this a safe way to keep secrets? There is something re-assuring to me that the difficulty of breaking a code depends not on obfuscation and misdirection, but on a deep mathematical problem that has eluded the talents of many of history's greatest minds (see, e.g., Gauss's quote above!). Other difficult mathematical problems are also used to hide secrets, some based on quite different questions, though one of my favorites is based on a problem that we now show to be just as difficult as factoring:

There are quick algorithms for taking square roots mod p when p is prime [4]. To extend these to composite n we need the factorization of n, since then we can find the square root modulo each of the prime power divisors of n, and then recover the square root mod n using the Chinese Remainder Theorem. Thus a fast factoring algorithm will give a fast algorithm for extracting square roots modulo composite integers n.

On the other hand suppose that we have a fast algorithm for extracting square roots modulo the composite integer n, and we wish to factor n. Then we can simply select a number $y \pmod n$ at random, feed the least residue of $y^2 \pmod n$ to our algorithm, which will return one of the square roots of $y^2 \pmod n$, say x. Then we have an at least 50% chance that $(x-y,n)$ $(x+y,n)$ provides a factorization of n. If we are unlucky we do this again and again until we succeed. The probability that we fail after doing this 100 times is negligible, no more than 1 in 2^{100}. Therefore we have proved that a fast algorithm for extracting square roots modulo composite integers n will give a fast factoring algorithm. Combining the last two paragraphs we see that these two problems are thus equally difficult.

7.3. **Primality testing.** *Fermat's little theorem* states that $a^p \equiv a \pmod p$ for every integer a whenever p is prime. Conversely if $a^n \not\equiv a \pmod n$ for some integer a then n is composite. One can compute $a^n \pmod n$ quite rapidly,[19] and so one can quickly prove that a given integer n is composite if $2^n \not\equiv 2 \pmod n$. It is perhaps a little surprising that this gives a proof that n is composite without producing any factors of n! If this test fails then we can check whether $3^n \not\equiv 3 \pmod n$, $5^n \not\equiv 5 \pmod n$,etc. Most composite n will be revealed in this way; if every composite were to be revealed in this way then the test would also serve as a primality test, primes being those numbers that are not revealed as composites. Unfortunately there are composite n for which $a^n \equiv a \pmod n$ for every integer a, for example 561 and 1729, and these *Carmichael numbers*, though rare, are infinite in number.

One can modify the above test through the following development of Fermat's little theorem: If $(a,p)=1$ then $a^{p-1} \equiv 1 \pmod p$, so that $a^{\frac{p-1}{2}} \equiv \pm 1 \pmod p$, as $a^{\frac{p-1}{2}}$ is the square root of a^{p-1}. If $a^{\frac{p-1}{2}} \equiv 1 \pmod p$ then we can take the square root again, and

[19] By the method of *successive squaring*.

again, up to as many as r times where 2^r divides $p-1$ but not 2^{r+1}. This sequence of residues must either be all 1's, or all 1's until we reach a -1. Anything else and we know that n is composite. There are no composite numbers n for which $a^{n-1} \pmod{n}$ and all its square roots have this property for all integers a which are coprime to n. In fact at least three-quarters of such values of a do not have this property if n is composite; these a are called *witnesses* to the compositeness of n. Thus we can distinguish primes from composites by looking for such witnesses, though there is no guaranteed quick way to find a witness. If we test 100 values of a chosen at random then the combined test will misidentify a composite number as prime, that is, it will fail to find a witness, with probability under 1 in 2^{200}, something that will never occur in practice. If we assume the Generalized Riemann Hypothesis we can prove that by picking just the first few values of a (up to $2(\log n)^2$) then we are guaranteed to find a witness for any composite n, and so we will have a true primality test.

However what was long wanted was a method that could be unconditionally proven to always work in a fast time. Such a test was found in 2002 by Agrawal, Kayal and Saxena (see [6]), based on the following theorem:

For given integer $n \geq 2$, let r be a positive integer $< n$, for which n has order $> (\log n)^2$ modulo r. Then n is prime if and only if

- n is not a perfect power,
- n does not have any prime factor $\leq r$, and
- $(x+a)^n \equiv x^n + a \mod (n, x^r - 1)$ for each integer $a, 1 \leq a \leq \sqrt{r}\log n$.

8. Further directions

In the absence of unique factorization one might desire a close analogue to the Euclidean algorithm: A ring of integers, R, in $\mathbb{Q}(\sqrt{d})$ is *Euclidean* if for any $\alpha, \beta \neq 0 \in R$ there exist $\gamma, \delta \in R$ such that $\alpha = \beta\gamma + \delta$ where $|\delta| < |\beta|$. See [9] for a charming discussion of this question.

Gauss explicitly showed how to "compose" two quadratic forms, the equivalent of multiplication of ideals in the appropriate quadratic field, which was written down explicitly by Dirichlet. Recently Bhargava came up with a fascinating new viewpoint on this: Consider eight integers $a_{i,j,k}, 0 \leq i, j, k \leq 1$. For $\ell = 1, 2, 3$, let M_ℓ, and N_ℓ, be the 2-by-2 matrices formed from the $a_{i,j,k}$ by letting the ℓth coordinate of the index equal 0, and 1, respectively. Let $f_\ell(x, y)$ be the determinant of the matrix $M_\ell x - N_\ell y$; then the quadratic forms f_1, f_2, f_3 all have the same discriminant and satisfy $f_1 f_2 f_3 = 1$ in the class group. See [2] and its sequels for extraordinary developments of these ideas.

9. Acknowledgements

Thanks to Jordan Bell and Henri Darmon for their careful reading of, and comments on, this article.

References

[1] Alan Baker, A concise introduction to the theory of numbers, Cambridge University Press, Cambridge, 1984.

[2] Manjul Bhargava, Higher composition laws I: A new view on Gauss composition and quadratic generalizations, *Annals Math.* **159** (2004), 217-250.

[3] Mary Joan Collison, The Unique Factorization theorem, *Math. Mag.* **53** (1980), 96-100.

[4] Richard Crandall and Carl Pomerance, Primes, a computational perspective, Springer, New York, 2005.

[5] Carl F. Gauss, Disquisitiones Arithmeticae, Leipzig, Fleischer, 1801.

[6] Andrew Granville, It is easy to determine whether a given integer is prime, *Bulletin of the American Mathematical Society*, **42** (2005) 3-38.

[7] G.H. Hardy and E.M. Wright, Introduction to the theory of numbers, Oxford University Press, New York, 1979.

[8] W. Knorr, Problems in the interpretation of Greek number theory: Euclid and the fundamental theorem, *Studies in the history and philosophy of science* **7** (1976) 353-368.

[9] H.W. Lenstra, Jr., Euclidean number fields, *The Mathematical Intelligencer*, I. **2** (1979) 6-15; II. **2** (1980) 73-77; III. **2** (1980) 99-103.

[10] H.W. Lenstra, Jr., Solving the Pell equation, *Notices of the American Mathematical Society*, **49** (2002) 182-192.

[11] Paulo Ribenboim, 13 lectures on Fermat's last theorem, Springer-Verlag, New York-Heidelberg, 1979.

[12] André Weil, Number theory: An approach through history from Hammurapi to Legendre, Birkhäuser, Boston, 1984.

Départment de Mathématiques et Statistique, Université de Montréal, CP 6128 succ Centre-Ville, Montréal, QC H3C 3J7, Canada

E-mail address: andrew@dms.umontreal.ca