Why is Algebra Important for Number Theory?

Robin Truax

March 2021

Contents

1 Motivation

This post seeks to answer a common question asked by students of abstract algebra: "why?" The abstraction may seem unncessary and obtuse. However, by giving an application to another, more naturally aesthetic area of math, we can help illuminate the power of this abstraction.

These notes are inspired by Math 154, taught by Brian Conrad, at Stanford.

2 Essential Ring Theory

In this section, we will review three important conditions on the structure of rings. These are likely familiar to someone with a course in algebra, but it's important to be familiar with the details to understand why these are such helpful notions. In these notes, all rings are commutative with identity.

We assume the knowledge of the following concepts from a first course on algebra: ring, subring, ideal, quotient, and unit. In these notes, R always denotes a ring and I always denotes an ideal.

2.1 Unique Factorization

Definition 1 (Irreducible)**.** A nonzero non-unit element $r \in R$ is called *irreducible* if there is no way to write $r = st$ with non-units s and t.

For example, 7 is irreducible in the ring of integers $\mathbb{Z}$.

Definition 2 (Division)**.** We say $r \in R$ *divides* t (and write $r \mid t$) if there exists $s \in R$ such that $rs = t$.

For example, 7 divides 14 in the ring of integers $\mathbb{Z}$, because $7 \cdot 2 = 14$.

Definition 3 (Prime)**.** A nonzero non-unit element $p \in R$ is called *prime* if, whenever $p \mid ab$, $p \mid a$ or $p \mid b$.

For example, 7 is a prime element in the ring of integers $\mathbb{Z}$.

In fact, with some investigation, one might discover that, in the integers, the prime elements and the irreducible elements are the same. Indeed, we will prove that $\mathbb{Z}$ is a unique factorization domain (UFD), which implies that the prime and irreducible elements are the same. However, unfortunately, this is not true in general. Let's take a look at an example.

Proposition 1. *Not all irreducible elements are prime in the ring* $\mathbb{Z}[\sqrt{-5}] = \{a + b\sqrt{-5} \mid a, b \in \mathbb{Z}\}$.

Proof. Notice that $2 \cdot 3 = 6$, so 2 divides 6. To see why 2 is irreducible, consider a map which describes the "size" of an element in $\mathbb{Z}[\sqrt{-5}]$, called the *norm map*:

$$N(a + b\sqrt{-5}) = a^2 + 5b^2.$$

It is not difficult to check via computation that $N(\alpha\beta) = N(\alpha)N(\beta)$ for any $\alpha, \beta \in \mathbb{Z}[\sqrt{-5}]$. Therefore, if $\alpha\beta = 2$, then $N(\alpha)N(\beta) = N(\alpha\beta) = N(2) = 4$. Either $N(\alpha)$ and $N(\beta)$ are both 2, or one of $N(\alpha)$ and $N(\beta)$ is 1, and the other is 4. Yet it is not difficult to check, from the formula, that there are no elements of norm 2. Therefore one of $N(\alpha)$ and $N(\beta)$ is 1. Yet the only elements with norm 1 are 1 and -1, which are both plainly units. Therefore, 2 cannot be written as the product of two non-units, so it is irreducible.

On the other hand, 6 is equal to $(1 + \sqrt{-5})(1 - \sqrt{-5})$. If 2 were prime, then it would divide at least one of these two elements. Yet it does not divide either, so it isn't prime. Thus, 2 is irreducible and not prime. □

Notice that the reason why this failed had something to do with the fact that we could write 6 as the product of irreducible elements in two different ways. Yet in $\mathbb{Z}$, each number can be uniquely expressed as the product of primes; we call this unique factorization. Therefore, we make the following definition.

Definition 4 (Unique Factorization Domain)**.** Let R be a ring. We say R is a *unique factorization domain* (UFD) if for any $r \in R$ we can write r as the product of irreducible elements $\pi_1, \ldots, \pi_n$ and a unit u, as so:

$$r = u\pi_1 \cdots \pi_n.$$

We also require this expression to be unique: if $r = v\rho_1 \cdots \rho_m$ is the product of a unit v and irreducible elements $\rho_1, \ldots, \rho_m$, $m = n$ and can rearrange the irreducible elements $\rho_1, \ldots, \rho_n$ such that π_i and ρ_i are associate (that is, $\pi_i = u_i\rho_i$ for a unit u_i) for each i.

Proposition 2. *In a UFD, every prime element is irreducible and vice versa.*

Proof. We'll leave this as an exercise, since it's not too hard. □

We'll see some examples later. In particular, we'll prove $\mathbb{Z}$ is a unique factorization domain, as we guessed.

2.2 Principal Ideals

Definition 5 (Principal Ideals)**.** Suppose I is an ideal of a ring R. Then, I is called *principal* if there exists a element $r \in R$ such that $I = (r)$; that is, I is the set of multiples of r.

Definition 6 (Principal Ideal Domain)**.** A ring R is called a *principal ideal domain* (PID) if every ideal of R is a principal ideal.

Why might we care about principal ideal domains? The following theorem provides one example.

Theorem 3. *Every principal ideal domain is a unique factorization domain.*

Proof. The proof of this theorem is not particularly illuminating, and it is quite laborious. Therefore, I cite this reference (clickable link) which explains the result in detail. □

2.3 Euclidean Domains

Definition 7 (Euclidean Domain)**.** A ring R is called a *Euclidean domain* if there exists a "norm map" $N : R \to \mathbb{Z}_{\geq 0}$ that satisfies the following requirements:

1. $N(r) = 0$ if and only if $r = 0$.
2. For any element $a \in R$ and nonzero $b \in R$, there exists $q, r \in R$ such that $N(r) < N(b)$

$$a = qb + r.$$

You might have an idea of why we care about Euclidean domains from the previous section.

Theorem 4. *Any Euclidean domain is a principal ideal domain.*

Proof. Consider an ideal $I \lhd R$. Let $b \in I$ be the nonzero element with the smallest norm of any nonzero element of I. I claim that $(b) = I$. To see why, take any $a \in I$. Then there are $q, r \in R$ such that

$$a = qb + r.$$

but $N(r) < N(b)$. Notice that $qb \in I$, so $r = a - qb \in I$. But this implies that $r = 0$, since b is has minimal norm of all nonzero elements of I. Therefore, a is a multiple of b, so (b) contains I. Since clearly I contains (b), since it contains b, we indeed have $I = (b)$, so I is principal, as desired. □

In particular, since $\mathbb{Z}$ is a Euclidean domain (look up, for example, the Euclidean algorithm), it is a principal ideal domain *and* a unique factorization domain. Here's another example:

Problem 1. Let $\mathbb{F}$ be a field, and $\mathbb{F}[X]$ be the ring of polynomials over said field. For each polynomial f, let $\deg f$ be the degree of f. By convention, the zero polynomial has degree $-\infty$. Prove that $\mathbb{F}[X]$ is a Euclidean domain with norm $N(f) = 2^{-\deg(f)}$. [Hint: the constant chosen – in this case 2 – doesn't matter.]

3 The Gaussian Integers

Next, we'll consider a few rings where are critical for number theory. We'll prove that they are unique factorization domains by finding a norm under which they are Euclidean domains.

The main ring which we are concerned with is called the *Gaussian integers*, and they are complex numbers of the form $a + bi$ where a and b are integers. The Gaussian integers, written $\mathbb{Z}[i]$ (which is pronounced "Z adjoin i"), has a close relationship to sums of squares.

3.1 $\mathbb{Z}[i]$ is a Euclidean Domain

Lemma 5. *The norm $N(x+yi) = x^2+y^2$ is multiplicative; that is, $N(\alpha\beta) = N(\alpha)N(\beta)$ for any $\alpha, \beta \in \mathbb{Z}[i]$.*

Proof. Left as an exercise in computation. □

Theorem 6. *$\mathbb{Z}[i]$ forms a Euclidean domain under the norm $N(x + yi) = x^2 + y^2$.*

Proof. Clearly, $N(a) = 0$ if and only if $a = 0$. Therefore, consider $a, b \in \mathbb{Z}[i]$ with b nonzero. Then, by multiplying the top and bottom of $\frac{a}{b}$ by the complex conjugate $\overline{b}$, we can write

$$\frac{a}{b} = t_1 + t_2 i$$

for rational numbers t_1 and t_2. Let q_1 and q_2 be the closest integers to t_1 and t_2, respectively, so that $q_1 + q_2 i$ is the Gaussian integer closest to $t_1 + t_2 i$. Then write $\varepsilon_1 = t_1 - q_1$ and $\varepsilon_2 = t_2 - q_2$, so that

$$\frac{a}{b} = t_1 + t_2 i = (q_1 + q_2 i) + (\varepsilon_1 + \varepsilon_2 i).$$

Abbreviate $q_1 + q_2 i$ by q, and $\varepsilon_1 + \varepsilon_2 i$ by ε. Then, $\frac{a}{b} = q + \varepsilon$ becomes $a = bq + b\varepsilon$. Define $r = b\varepsilon$, so that $a = bq + r$. Now, all we need to prove is that $N(r) < N(b)$. To do this, notice that $N(r) = N(b\varepsilon) = N(b)N(\varepsilon)$, so the key is to prove that $N(\varepsilon) < 1$.

This follows from one incredibly important fact: $|\varepsilon_1| \leq \frac{1}{2}$ and $|\varepsilon_2| \leq \frac{1}{2}$. This is because any rational number is at most a half-unit away from some integer. As a result,

$$N(\varepsilon) = N(\varepsilon_1 + \varepsilon_2 i) \leq \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} < 1.$$

Therefore, $N(r) = N(b\varepsilon) = N(b)N(\varepsilon) < N(b)$, exactly as desired! Notice that $N(b) \neq 0$, because $b \neq 0$. □

Corollary 6.1. *$\mathbb{Z}[i]$ is a principal ideal domain and unique factorization domain.*

3.2 The Power of Abstraction: $\mathbb{Z}[\sqrt{-2}]$

It might seem like all of this is abstraction for abstraction's sake, but notice that the earlier argument ports over very well to another subring of $\mathbb{C}$: $\mathbb{Z}[\sqrt{-2}] = \{a + b\sqrt{-2} \mid a, b \in \mathbb{Z}\}$. In fact,

Lemma 7. *The norm $N(x+y\sqrt{-2}) = x^2+2y^2$ is multiplicative; that is, $N(\alpha\beta) = N(\alpha\beta)$ for any $\alpha, \beta \in \mathbb{Z}[i]$.*

Theorem 8. *$\mathbb{Z}[\sqrt{-2}]$ forms a Euclidean domain under the norm $N(x + y\sqrt{-2}) = x^2 + 2y^2$.*

Proof. The proof of this fact is exactly the same as for $\mathbb{Z}[i]$, only now computing $N(\varepsilon)$ is slightly different:

$$N(\varepsilon) = N(\varepsilon_1 + \varepsilon_2 i) \leq \left(\frac{1}{2}\right)^2 + 2\left(\frac{1}{2}\right)^2 = \frac{1}{4} + 2 \cdot \frac{1}{4} = \frac{3}{4} < 1.$$

□

4 Diophantine Equations

Finally, once we've developed this algebra, we can begin applying it to *Diophantine equations*, central objects of study in number theory. Diophantine equations ask us to find *integer* solutions, which are often significantly harder than finding real or complex solutions. We will grapple with a few famous examples in the coming sections, relying on the structure of the rings studied earlier.

As for prerequisites, we will use the basics of modular arithmetic (including, for example, that $\mathbb{Z}/p\mathbb{Z}$ is a field) and two facts from elementary number theory called *supplements to the law of quadratic reciprocity.* These facts have elementary proofs, and are as follows:

Lemma 9. *If p is an odd prime, there exists x such that $x^2 \equiv -1 \bmod p$ if and only if $p \equiv 1 \bmod 4$.*

Lemma 10. *If p is an odd prime, there exists x such that $x^2 \equiv 2 \bmod p$ if and only if $p \equiv 1$ or $7 \bmod 8$.*

4.1 Which Integers are Sums of Squares? Ask $\mathbb{Z}[i]$.

This section concerns itself with discovering which positive integers n can be expressed as the sum of two perfect squares. For example, $17 = 4^2 + 1^2$ but 33 cannot be written as the sum of two perfect squares. Following are our first two hints that Gaussian integers might be the key to the puzzle of sums of squares.

Lemma 11. *A positive integer n can be expressed as the sum of two perfect squares if and only if there exists $\alpha \in \mathbb{Z}[i]$ with $N(\alpha) = n$.*

Proof. Let $\alpha = x + yi$ be such that $N(\alpha) = n$. Then $n = N(\alpha) = x^2 + y^2$. On the other hand, if $n = x^2 + y^2$, then assigning $\alpha = x + yi$ we have $N(\alpha) = x^2 + y^2 = n$. □

Lemma 12. *If m and n can be expressed as the sum of two squares, then mn can be expressed as the sum of two squares.*

Proof. By Lemma 11, $m = N(\alpha)$ and $n = N(\beta)$ for Gaussian integers α and β. But the norm is multiplicative, so $mn = N(\alpha)N(\beta) = N(\alpha\beta)$. Thus, again by Lemma 11, mn can be expressed as the sum of two squares. □

Theorem 13. *An odd prime p can be expressed as the sum of two squares iff -1 is a square mod p.*

Proof. If $p = x^2 + y^2$, then $x^2 + y^2 \equiv 0 \bmod p$. Then, $x^2 \equiv -y^2 \bmod p$. Now, neither of x or y can be divisible by p (since then $x^2 + y^2$ would be at least p^2, which is too large), so we may divide by y^2 to see

$$\left(\frac{x}{y}\right)^2 \equiv -1 \bmod p.$$

On the other hand, if -1 is a square mod p, then by definition there exists n such that $n^2 \equiv -1 \bmod p$. Then $n^2 + 1 \equiv 0 \bmod p$. In other words, p divides $n^2 + 1$. Now it's time to enter the world of Gaussian integers:

$$n^2 + 1 = (n + i)(n - i).$$

Now, p divides n^2+1, so it divides $(n+i)(n-i)$. But $n+i$ is not a multiple of p, because if $n+i = p(a+bi) = pa+pbi$, then by comparing coefficients $pb = 1$, which is impossible. Similar logic implies that $n - i$ is not a multiple of p. But, if p divides $(n + i)(n - i)$ but not $n + i$ or $n - i$, it cannot be prime. Therefore, since in every UFD, prime and irreducible elements are the same, p is not irreducible.

This means that $p = \alpha\beta$ for some nonzero non-units α and β. But this implies $p^2 = N(p) = N(\alpha\beta) = N(\alpha)N(\beta)$. Now, the only elements of norm 1 are 1 and -1, both units, so α and β both have norm greater than 1. Therefore, $N(\alpha) = N(\beta) = p$, so p can be expressed as the sum of two squares by Lemma 11. □

Corollary 13.1. *A prime p can be expressed as the sum of two squares if and only if $p = 2$ or $p \equiv 1 \bmod 4$.*

Proof. Clearly, if $p = 2$, then $1^2 + 1^2 = p$. The rest follows using Lemma 9 on the above theorem. □

Lemma 14. *Every irreducible $\pi \in \mathbb{Z}[i]$ either has norm 2, $p \equiv 1 \bmod 4$ or p^2 for an odd prime p.*

Proof. Suppose that $N(\pi)$ has prime factorization $p_1 \cdots p_m$. Now, $N(\pi) = \pi\overline{\pi}$, where $\overline{\pi}$ is the complex conjugate of π. Therefore, π divides $p_1 \cdots p_m$. Now, π is irreducible so it is prime since $\mathbb{Z}[i]$ is a UFD. Thus, π divides p_i for some i. By the multiplicativity of the norm, $N(\pi)$ divides $N(p_i) = p_i^2$. Therefore, since $N(p_i)$ cannot equal 1, either $N(\pi) = p_i$ or $N(\pi) = p_i^2$. Yet the former case is only possible if $p_i = 2$ or $p_i \equiv 1 \bmod 4$, and this implies the desired result. □

Theorem 15. *A positive integer n can be expressed as the sum of two squares if and only if every prime of the form $p \equiv 3 \bmod 4$ appears to an even power in the prime factorization of n.*

Proof. Suppose every prime of the form $p \equiv 3 \bmod 4$ appears to an even power in the prime factorization of n. Then, n can be written as the product of some number of 2s, primes of the form $p \equiv 1 \bmod 4$, and p^2 for $p \equiv 3 \bmod 4$. Now, each of these can be written as the sum of two squares, either by our previous theorem or by $p^2 = p^2 + 0^2$. Therefore, by Lemma 12, n is the sum of two squares.

On the other hand, suppose n is the sum of two squares. Then $n = N(\alpha)$ for $\alpha \in \mathbb{Z}[i]$. Factor α into irreducibles $\pi_1 \cdots \pi_m$, so that

$$n = N(\alpha) = N(\pi_1) \cdots N(\pi_m).$$

By Lemma 14, each of these norms are either 2, a prime $p \equiv 1 \bmod 4$, or p^2 for an odd prime p. But this expression demonstrates that each prime power $p \equiv 3 \bmod 4$ can only appear in square terms; i.e., to an even power. Hence we are done. □

4.2 Abstraction Strikes Again: $\mathbb{Z}[\sqrt{-2}]$ and $n = x^2 + 2y^2$

Suppose we are trying to find which positive integers can be expressed in the form $x^2 + 2y^2$. It turns out that we can analogize all of our arguments from the previous section to this problem! In particular, instead of working over $\mathbb{Z}[i]$, we work over $\mathbb{Z}[\sqrt{-2}]$. I state the analogous lemmas and theorems here, and suggest as an exercise that you adapt our previous arguments.

Lemma 16. *A positive integer n can be expressed in the form x^2+2y^2 if and only if there exists $\alpha \in \mathbb{Z}[\sqrt{-2}]$ with $N(\alpha) = n$.*

Lemma 17. *If m and n can be expressed in the form $x^2 + 2y^2$, then mn can be expressed in said form.*

Theorem 18. *An odd prime p can be expressed in the form $x^2 + 2y^2$ iff 2 is a square mod p.*

Corollary 18.1. *A prime p can be expressed in the form $x^2 + 2y^2$ if and only if $p = 2$ or $p \equiv 1$ or $7 \bmod 8$.*

Lemma 19. *Every irreducible $\pi \in \mathbb{Z}[\sqrt{-2}]$ either has norm 2, $p \equiv 1$ or $7 \bmod 8$, or p^2 for an odd prime p.*

Theorem 20. *A positive integer n can be expressed as the sum of two squares if and only if every prime of the form $p \equiv 3, 5 \bmod 8$ appears to an even power in the prime factorization of n.*

4.3 Integer Solutions to the Elliptic Curve $y^2 = x^3 - 2$

Next, we will use $\mathbb{Z}[\sqrt{-2}]$ to approach another type of problem entirely: finding all integer solutions to the elliptic curve $y^2 = x^3 - 2$. First, rearrange the curve into the form $y^2 + 2 = x^3$. Then, notice that y cannot be even. This is because if y is even, then $x^3 = y^2 + 2$ must be even, whence x is even. But then y^2 and x^3 are both 0 mod 4, which is impossible since $x^3 \equiv y^2+2 \equiv 2 \bmod 4$ (if y is even). Therefore, y is odd.

Recall that we have rearranged the curve into the form $y^2 + 2 = x^3$. In $\mathbb{Z}[\sqrt{-2}]$, this can be factored as

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

This may seem useless, but there is an important fact at play: $(y+\sqrt{-2})$ and $(y-\sqrt{-2})$ are coprime. More precisely, they share no non-unit factors. To see why, suppose that d is a non-unit which divides $(y-\sqrt{-2})$ and $(y+\sqrt{-2})$. Then d divides their difference, $-2\sqrt{-2} = (\sqrt{-2})^3$. Since $\mathbb{Z}[\sqrt{-2}]$ is a UFD, this implies that $\sqrt{-2}$ divides d, and therefore that $\sqrt{-2}$ divides $(y - \sqrt{-2})$ and $(y + \sqrt{-2})$. But this is simply untrue:

$$(a + b\sqrt{-2})\sqrt{-2} = -2b + a\sqrt{-2}.$$

Since y cannot be even, the right-hand side cannot equal $y + \sqrt{-2}$ or $y - \sqrt{-2}$ for any choice of a or b. Therefore, we have found a contradiction, and no non-unit divides $(y + \sqrt{-2})$ and $(y - \sqrt{-2})$.

Now, $(y + \sqrt{-2})$ and $(y - \sqrt{-2})$ are two coprime numbers which multiply to make a perfect cube. Yet, if two coprime numbers multiply to make a perfect cube, what may we conclude about the original numbers? Well, that they themselves are perfect cubes! Therefore, $(y + \sqrt{-2})$ and $(y - \sqrt{-2})$ are perfect cubes. From here, it is only a matter of computation: if $(a + b\sqrt{-2})^3 = (y + \sqrt{-2})$ then by expanding we have

$$a(a^2 - 6b^2) + b(3a^2 - 2b^2)\sqrt{-2} = y + \sqrt{-2}.$$

Comparing the coefficients of $\sqrt{-2}$, we have $b(3a^2 - 2b^2) = 1$. There are two cases to investigate:

Case 1: $b = 3a^2 - 2b^2 = -1$.
Notice that, mod 3, the equation $3a^2 - 2b^2 = -1$ becomes $-2b^2 \equiv -1 \bmod 3$, which is equivalent to $b^2 \equiv -1 \bmod 3$. Since -1 is not a quadratic residue mod 3, this has no solutions.

Case 2: $b = 3a^2 - 2b^2 = 1$.
In this case, $3a^2 = 3$, whence $3a^2 = 3 \Rightarrow a = \pm 1$. Now it is only a matter of backtracking: again, by comparing coefficients, $a(a^2 - 6b^2) = y$, whence $y = \pm 5$. From here, it is trivial to show that if $y = \pm 5$, then $x = 3$. Therefore, the only integer solutions (x, y) to the elliptic curve are $(3, 5)$ and $(3, -5)$.