154: Diophantine Equations

Nir Elber

Fall 2023

Contents

How strange to actually have to see the path of your journey in order to make it.

—Neal Shusterman, [Shu16]

Theme 1
Linear Equations

Think deeply of simple things

—Ross Program, [Pro22]

1.1 Modular Arithmetic and Sage

In this section, we review the elementary number theory we will use in these notes. The goal of the present chapter is to be able to solve the equation

$$ax + by = 1$$

as quickly as possible, but we will encounter Diophantine approximation in the process.

1.1.1 Local Obstructions

A theme that will reappear in this course is that of "local obstructions," so we introduce the idea now. Here are some examples.

Example 1.1. The only integer solution to the equation $x^2 + y^2 = 3z^2$ is $(x, y, z) = (0, 0, 0)$.

Solution. Of course $(0, 0, 0)$ is a solution, so the main content is showing that it is the only one. Suppose that (x, y, z) is a nonzero solution, and we suppose that (x, y, z) is minimal with respect to $|x| + |y| + |z| > 0$. If all the terms are even, then $(x/2, y/2, z/2)$ is also an integer solution with $|x/2| + |y/2| + |z/2| < |x| + |y| + |z|$, violating minimality. Thus, we may assume that at least one of the terms is odd. We have two cases; the main point is that $x^2 \equiv 0, 1 \pmod 4$ for any integer x.

- If z is odd, then we are asking for
$$x^2 + y^2 \equiv 3 \pmod 4.$$
But $x^2, y^2 \pmod 4 \in \{0, 1\}$ cannot achieve this.
- If z is even, then we are asking for
$$x^2 + y^2 \equiv 0 \pmod 4.$$
However, without loss of generality we will have x odd and so $x^2 \equiv 1 \pmod 4$. But then $x^2+y^2 \equiv 1+y^2 \pmod 4$ will never be $0 \pmod 4$.

All cases have caused contradiction, so we have finished the proof. ■

Example 1.2. There are no integer solutions to the equation $6x + 9y = 2$.

Solution. Reducing $\pmod 3$ means that any integer solution to $6x + 9y = 2$ implies $0 \equiv 2 \pmod 3$, which is a contradiction. ■

Now that we've seen some examples, let's make explicit what is going on.

Idea 1.3. Given an equation $f(x_1, \ldots, x_n) = 0$, we can check if f has solutions in $\mathbb{Z}$ by first checking if there are solutions to

$$f(x_1, \ldots, x_n) \equiv 0 \pmod m$$

for integers m.

What is useful about Idea 1.3 is that checking for solutions $\pmod m$ amounts to a finite computation where variables live in $\mathbb{Z}/m\mathbb{Z}$, and we can simply run the finite computation to check.

Of course, Idea 1.3 is not perfectly robust, but it will guide our discussion of Diophantine equations throughout this course.

Non-Example 1.4. One can show that

$$(x^2 - 2)(x^2 - 3)(x^2 - 6) = 0$$

has solutions $\pmod p$ for all primes p, but there is no integer solution.

Here is an example which is akin to Idea 1.3 but not quite the same.

Example 1.5. There are no integer solutions to $x^2 + y^2 = 2xy - 1$.

Solution. This equation is actually $(x - y)^2 = -1$, which has no solutions because $(x - y)^2 > -1$ for any real numbers $x, y \in \mathbb{R}$. ■

Example 1.6. There are no integer solutions to $x^2 + y^2 = 6$.

Solution. We see that $x \in \{0, \pm 1, \pm 2\}$ forces $y \in \{\pm\sqrt{6}, \pm\sqrt{5}, \pm\sqrt{2}\}$, none of which provide integer solutions. However, if $|x| \geq 3$, then

$$x^2 + y^2 = 9 + y^2 > 6,$$

from which we see that there are not even real solutions! ■

The above examples teach us that it is also useful to check for real-valued solutions to an equation in addition to checking $\pmod m$ for various integers m. These are also "local obstructions."

1.1.2 The Law of Linear Reciprocity

Idea 1.3 is useful for determining when a linear equation of the form $ax + by = 1$ cannot have solutions. The goal of the present section is to show that these "local obstructions" are the only obstructions. Namely, we will prove a result of the following type.

Proposition 1.7. Let a, b, and c be integers. Then there are integers $x, y \in \mathbb{Z}$ such that $ax + by = c$ if and only if, for any integer m, there are integers $x_m, y_m \in \mathbb{Z}$ such that

$$ax_m + by_m \equiv c \pmod{m}.$$

In other words, it is enough to check locally. However, Proposition 1.7 is not very helpful for actually trying to determine if $ax + by = c$ has solutions: we would have to check $ax + by \equiv c \pmod{m}$ for infinitely many moduli m, which is not a finite computation! Thankfully, we have the following more effective version of Proposition 1.7.

Proposition 1.8. Let a, b, and c be integers. Then there are integers $x, y \in \mathbb{Z}$ such that $ax + by = c$ if and only if there are integers $x, y \in \mathbb{Z}$ such that

$$ax + by \equiv c \pmod{b}.$$

In other words, the only modulus we have to check is $m = b$. Let's prove Proposition 1.8.

Proof of Proposition 1.8. Of course having integers x and y such that $ax + by = c$ will imply that $ax + by \equiv c \pmod{b}$. Conversely, suppose we have integers x_0 and y_0 such that

$$ax_0 + by_0 \equiv c \pmod{b}.$$

Then we know there is some integer y_1 such that

$$ax_0 + by_0 = c + by_1,$$

so $ax_0 + b(y_0 - y_1) = c$ provides an integer solution to $ax + by = c$. ■

Example 1.9. The equation $3x + 5y = 1$ has integer solutions.

Solution. By Proposition 1.8, it suffices to check $\pmod{3}$. Then we are looking for integers x and y such that

$$3x + 5y \equiv 1 \pmod{3}.$$

Well, $(x, y) = (0, 2)$ will do the trick. ■

Example 1.10. The equation $2x + 4y = 3$ has no integer solutions.

Solution. By Proposition 1.8, it suffices to check $\pmod{2}$. Then we are looking for integers x and y such that

$$2x + 4y \equiv 3 \pmod{2}.$$

But this implies $0 \equiv 3 \pmod{2}$, which is a contradiction, so there can be no integer solutions. ■

Proposition 1.8 also allows us to prove the "reciprocity" theorem. These are also a major theme in number theory, though we will not see even close to the full story in this course. What is remarkable in the following result is that we have found a way to switch the modulus of our "local obstruction" around, perhaps at the cost of adjusting the equation being considered. Such statements are in general very profitable!

Proposition 1.11 (law of linear reciprocity)**.** Let a, b, and c be integers. Then there is an integer x such that $ax \equiv c \pmod{b}$ if and only if there is an integer x such that $bx \equiv c \pmod{a}$.

Proof. There is an integer x such that $ax \equiv c \pmod{b}$ if and only if there are integers x and y such that $ax = c - by$, which is equivalent to

$$ax + by = c.$$

This condition is now symmetric in a and b, so running the above argument backwards provides equivalence to finding an integer x such that $bx \equiv c \pmod{a}$. ■

Example 1.12. The equation $93x + 35y = 1$ has integer solutions.

Solution. By Proposition 1.8, it is equivalent to check that

$$23x \equiv 93x + 35y \equiv 1 \pmod{35}$$

has integer solutions. By Proposition 1.11, this is equivalent to having integer solutions to

$$12x \equiv 35x \equiv 1 \pmod{23}.$$

Going again, by Proposition 1.11, this is equivalent to having integer solutions to

$$11x \equiv 23x \equiv 1 \pmod{12}.$$

Continuing, by Proposition 1.11, this is equivalent to having integer solutions to

$$x \equiv 12x \equiv 1 \pmod{11},$$

for which we see that $x = 1$ works. ■

Example 1.13. The equation $289x + 323y = 2$ has no integer solutions.

Solution. By Proposition 1.8, it is equivalent to check that

$$34y \equiv 289x + 323y \equiv 2 \pmod{289}$$

has integer solutions. By Proposition 1.11, this is equivalent to having integer solutions to

$$17x \equiv 289x \equiv 2 \pmod{34}.$$

One more time, Proposition 1.11 says that it is equivalent to have integer solutions to

$$0 \equiv 34x \equiv 2 \pmod{17},$$

which is false. ■

1.1.3 Bézout's Theorem

Proposition 1.11 does a good job of determining when there are integer solutions to an equation of the form $ax + by = c$, but we would like a more efficient characterization, and we would also like an efficient way to write down the solutions. We begin with the more uniform characterization.

Theorem 1.14 (Bézout)**.** Let a, b, and c be integers. Then there are integers x and y such that $ax + by = c$ if and only if $\gcd(a, b)$ divides c.

We are going to prove Theorem 1.14 multiple times, essentially to emphasize different points of view on this area of number theory. To begin, let's establish that Proposition 1.11 is in fact able to provide a proof.

Proof of Theorem 1.14 via Proposition 1.11. We imitate the previous examples. Note that $ax + by = c$ if and only if $(-a)(-x) + by = c$ and similar for other choices of signs, so we might as well assume that a and b and c are all nonnegative integers. Additionally, having solutions for $ax + by = c$ is a condition symmetric on a and b, so we might as well assume that $a \leq b$.

We induct on a. If $a = 0$, then either $b = 0$, and we have a solution if and only if $c = 0 = \gcd(a, b)$, or $b \neq 0$, and we have a solution if and only if $c = by = \gcd(a, b)y$ for some integer y. Otherwise, $a > 0$. Now, by Proposition 1.8, we have an integer solution if and only if

$$ry \equiv ax + by \equiv c \pmod{a}$$

has an integer solution, where r is chosen so that $b \equiv r \pmod{a}$ and $0 \leq r < a$. By Proposition 1.11, this is now equivalent to having an integer solution to

$$ax \equiv c \pmod{b - a},$$

which by Proposition 1.8 is equivalent to having an integer solution to $rx+ay = c$. But now we have replaced (a, b) with (r, a), where $r < a$ and $\gcd(a, b) = \gcd(r, a)$, so induction completes the argument. ■

The above argument is fairly involved, so it is rewarding to know that the following cleaner proof exists.

Proof of Theorem 1.14 via well-ordering. It suffices to show that

$$\{ax + by : x, y \in \mathbb{Z}\} = \gcd(a, b)\mathbb{Z}.$$

Quickly, if $a = b = 0$, then both sides are $\{0\}$, so there is nothing to say. Otherwise, we may assume that at least one of a or b is nonzero. Certainly $\gcd(a, b)$ divides $ax + by$ for any $x, y \in \mathbb{Z}$, so $\{ax + by : x, y \in \mathbb{Z}\} \subseteq \gcd(a, b)\mathbb{Z}$. It remains to show the other inclusion, which is equivalent to showing $\gcd(a, b) \in \{ax+by : x, y \in \mathbb{Z}\}$.

Well, we expect $\gcd(a, b)$ to be the smallest positive element of $\{ax + by : x, y \in \mathbb{Z}\}$, so we let g denote this smallest positive element, and we want to show that $g = \gcd(a, b)$. (This g exists by the well-ordering of $\mathbb{N}$. Note that $\{ax + by : x, y \in \mathbb{Z}\}$ certainly has some positive element because it contains $a^2 + b^2 > 0$.) Certainly $\gcd(a, b)$ divides g by the argument of the previous paragraph, so it suffices to show that g divides $\gcd(a, b)$, for which we will show that $g \mid a$ and $g \mid b$.

In fact, we will only show that $g \mid a$, and $g \mid b$ follows symmetrically. For this, we use the division algorithm to write

$$a = gq + r$$

for some integers $q, r \in \mathbb{Z}$ where $0 \leq r < g$. Now, $r = a - gq$ will live in $\{ax + by : x, y \in \mathbb{Z}\}$, but $r < g$ forces r to not be a positive element in this set by minimality, so we must have $r = 0$. Thus, $a = gq$, which means $g \mid a$, as needed. ■

The drawback of the above cleaner proof is that it is difficult to see how to turn it into an effective algorithm to actually compute x and y. Indeed, the argument does not even make it clear how to find $x, y \in \mathbb{Z}$ such that

$$ax + by = \gcd(a, b),$$

which is in some sense the crux of the matter because we can then multiply x and y by $c/\gcd(a, b)$. With some care, we will be able to provide an effective algorithm, but it will take some care.

1.1.4 The Extended Euclidean Algorithm

The motivation to our algorithm will begin with wanting an efficient way to compute $\gcd(a, b)$, which we need to use Theorem 1.14 anyway. The Euclidean algorithm is based on the following lemma.

Lemma 1.15. Let a and b be integers. For any integer q, we have $\gcd(a, b) = \gcd(a - bq, b)$.

Proof. Note that an integer d divides a and b implies that d divides $a - bq$ and b; the converse holds by a symmetric argument. Thus, the conclusion follows from taking the least elements of the sets

$$\{d \in \mathbb{Z}_{\geq 0} : d \mid a \text{ and } d \mid b\} = \{d \in \mathbb{Z}_{\geq 0} : d \mid a - bq \text{ and } d \mid b\},$$

finishing. ■

We are now equipped to see an example of the Euclidean algorithm.

Example 1.16. We use the "Euclidean algorithm" to compute $\gcd(93, 35)$.

Solution. To begin, we repeatedly use the division algorithm to write

$$\begin{aligned} 93 &= 2 \cdot 35 + 23 \\ 35 &= 1 \cdot 23 + 12 \\ 23 &= 1 \cdot 12 + 11 \\ 12 &= 1 \cdot 11 + 1 \\ 11 &= 11 \cdot 1 + 0. \end{aligned}$$

Thus, repeatedly applying Lemma 1.15, we see

$$\gcd(93, 35) = \gcd(35, 23) = \gcd(23, 12) = \gcd(12, 11) = \gcd(11, 1) = 1,$$

which is what we wanted. ■

Exercise 1.17. Use the Euclidean algorithm to compute $\gcd(47, 31)$.

It is somewhat technical to make a rigorous argument avoid the above process. Take a moment to read and digest the following statement.

Proposition 1.18 (Euclidean algorithm)**.** Let a_0 and a_1 be positive coprime integers. Define the integer sequences $a_2, a_3, \ldots$ and $q_0, q_1, \ldots$ recursively by

$$a_n = q_n a_{n+1} + a_{n+2} \qquad \text{where} \qquad 0 \leq a_{n+2} < a_{n+1}$$

where $q_n \coloneqq \lfloor a_n / a_{n+1} \rfloor$ if $a_{n+1} > 0$ and $(a_{n+2}, q_n) \coloneqq (0, 0)$ otherwise. Then there is a minimal N such that $a_n = 0$ for $n > N$, and $a_N = \gcd(a_0, a_1)$.

Proof. By construction of the sequence, if $a_{n+1} > 0$, then $0 \leq a_{n+2} < a_{n+1}$. Thus, if $a_{n+1} > 0$ always, then $a_1, a_2, \ldots$ is a strictly decreasing sequence of positive integers, which is impossible by the well-ordering of the positive integers.

So indeed, there is some integer N such that $a_{N+1} = 0$, and we may choose N to be minimal with this property so that $a_N \neq 0$. (Note that $a_0 \neq 0$, so there is some n with $a_n \neq 0$.) Then $a_{N+1} = 0$ by construction, and the definition of our recursion enforces $a_n = 0$ for all $n > N$.

It remains to show that $a_N = \gcd(a_0, a_1)$. The main claim is that $\gcd(a_0, a_1) = \gcd(a_n, a_{n+1})$ for any $0 \leq n \leq N$, which will complete the proof by plugging in $n = N$. We show this claim by induction: there is nothing to say for $n = 0$, and for any $n < N$ so that $a_{n+1} > 0$, we see that

$$\gcd(a_n, a_{n+1}) = \gcd(q_n a_{n+1} + a_{n+2}, a_{n+1}) = \gcd(a_{n+1}, a_{n+2}),$$

which completes the inductive step. ■

Proposition 1.18 grants us another proof of Theorem 1.14.

Proof of Theorem 1.14 via Proposition 1.18. As usual, we start off with the "easier" direction: if $ax+by=c$ for some $x, y \in \mathbb{Z}$, then we note $\gcd(a,b)$ divides $ax+by$ and so divides c.

We use Proposition 1.18 to show the harder direction. Both the condition $ax+by=c$ and $\gcd(a,b) \mid c$ remain invariant to adjusting the sign of a and b, so we may assume $a, b \geq 0$. Additionally, if $a=0$, then both conditions are equivalent to $b \mid c$; a symmetric argument works for $b=0$. Thus, we may assume that $a, b > 0$.

Now, set $a_0 \coloneqq a$ and $a_1 \coloneqq b$ and build the sequence $a_2, a_3, \ldots$ of Proposition 1.18. By induction, we see that

$$a_n \in \{a_0 x + a_1 y : x, y \in \mathbb{Z}\}.$$

Indeed, there is nothing to say for $n=0$ and $n=1$. Then for the induction, we note that $\{a_0x+a_1y : x, y \in \mathbb{Z}\}$ is closed under $\mathbb{Z}$-linear combination, so containing a_n and a_{n+1} implies containing $a_{n+2} = a_n - q_n a_{n+1}$. Thus, using Proposition 1.18, we see that $a_N = \gcd(a,b)$ takes the form $ax+by$ for $x, y \in \mathbb{Z}$, completing the proof. ■

We are finally able to read the above proof closely to have an effective algorithm to compute x and y solving $ax+by=\gcd(a,b)$. This is called the "extended Euclidean algorithm" and is best seen by example.

Example 1.19. We use the "extended Euclidean algorithm" to find integers x and y such that $93x+35y=1$.

Proof. The idea is to run the Euclidean algorithm backwards "solving" for the remainders. Indeed, using the computations of Example 1.16, we see

$$\begin{aligned} 1 &= 12 - 1 \cdot 11 \\ 11 &= 23 - 1 \cdot 12 \\ 12 &= 35 - 1 \cdot 23 \\ 23 &= 93 - 2 \cdot 35. \end{aligned}$$

We now plug in for each successive remainder, writing

$$\begin{aligned} 1 &= 12 - 1 \cdot 11 \\ &= 12 - 1 \cdot (23 - 1 \cdot 12) = 2 \cdot 12 - 1 \cdot 23 \\ &= 2 \cdot (35 - 1 \cdot 23) - 1 \cdot 23 = 2 \cdot 35 - 3 \cdot 23 \\ &= 2 \cdot 35 - 3 \cdot (93 - 2 \cdot 35) = 8 \cdot 35 - 3 \cdot 93. \end{aligned}$$

Thus, $(x,y) = (-3, 8)$ will do the trick. ■

Exercise 1.20. Use the extended Euclidean algorithm to find integers x and y such that $47x+31y=1$.

1.1.5 Problems

Do at least ten points worth of the following exercises.

Problem 1.1.1 (1 point). Let $n \equiv 3 \pmod 4$. Show that there are not two integers $x, y \in \mathbb{Z}$ such that $x^2+y^2=n$.

Problem 1.1.2 (2 points). Let $n \equiv 7 \pmod 8$. Show that there are not three integers $x, y, z \in \mathbb{Z}$ such that $x^2+y^2+z^2=n$.

Problem 1.1.3 (2 points). Let a and b be integers. Suppose that there are pairs of integers (x, y) and (x', y') such that $ax + by = ax' + by' = 1$. Show that

$$x \equiv x' \pmod{b} \qquad \text{and} \qquad y \equiv y' \pmod{a}.$$

Problem 1.1.4 (2 points). Define the Fibonacci sequence $\{F_n\}_{n=0}^{\infty}$ by $F_0 = 0$, $F_1 = 1$, and $F_{n+2} = F_{n+1} + F_n$ for any $n \geq 0$. Show that $\gcd(F_{n+1}, F_n) = 1$ for any $n \geq 0$.

Problem 1.1.5 (3 points). Compute $\gcd(1027, 1738)$. Then find integers x and y such that $1027x + 1738y = \gcd(1027, 1738)$.

Problem 1.1.6 (3 points). Let a, b, and c be integers with $\gcd(a, b, c) = 1$. Show that there exist integers $x, y, z \in \mathbb{Z}$ such that $ax + by + cz = 1$.

Problem 1.1.7 (5 or 6 points). Implement the extended Euclidean algorithm.

(a) For five points, write (and submit) a function in Python with takes as input two coprime positive integers a and b and outputs integers x and y such that $ax + by = 1$. Your function should implement the extended Euclidean algorithm.

(b) For an additional point, make the function work for any coprime integers a and b.

Your test case is $(a, b) = (12345678901, 10987654321)$.

1.2 Finite Continued Fractions

In this section, we begin our discussion of continued fractions with a discussion of finite continued fractions. The reward for our efforts will be a more memory-efficient version of the extended Euclidean algorithm.

1.2.1 Connection to Continued Fractions

We begin with the definition of a continued fraction.

Definition 1.21 (continued fraction). A *continued fraction* expansion is an expression of the form

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \ddots}}},$$

which we will notate by $[a_0; a_1, a_2, \ldots]$. The terms $a_\bullet$ are the *continued fraction coefficients*.

In our application, the terms $a_0, a_1, a_2, \ldots$ will always be integers, and $a_1, a_2, \ldots$ will always be positive integers, but we take the moment to remark that this definition operates just fine even if these are not integers. This specialization does guarantee that we never run into division-by-zero problems, which is its principal advantage.

Remark 1.22. For the present section, our continued fractions will always be finite in length. In other words, our continued fractions will look like $[a_0; a_1, a_2, \ldots, a_n]$ for some perhaps large n. In the next section, we will allow continued fractions to have infinite length by defining

$$[a_0; a_1, a_2, \ldots] \coloneqq \lim_{n\to\infty} [a_0; a_1, a_2, \ldots, a_n],$$

but we will have to prove that this limit exists before providing this definition.

Continued fractions will be very interesting to us in the sequel, approximately speaking because they provide good rational approximations to real numbers. To start us off, suppose we have a real number α, and we would like to find coefficients $a_0, a_1, a_2, \ldots \in \mathbb{Z}$ such that $\alpha = [a_0; a_1, a_2, \ldots]$. In fact, we will be able to enforce $a_1, a_2, \ldots \in \mathbb{Z}_{\geq 0}$. To see how, note that if we want

$$\alpha = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \ddots}}},$$

then we should have $a_0 \coloneqq \lfloor \alpha \rfloor$. Once we agree what a_0 should be, we may rearrange this equation into

$$\frac{1}{\alpha - a_0} = a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \ddots}}.$$

Now we are trying to compute the continued fraction for $(\alpha - \lfloor \alpha \rfloor)^{-1} > 1$, so we may recurse. Namely, set $a_1 \coloneqq \left\lfloor (\alpha - \lfloor \alpha \rfloor)^{-1} \right\rfloor$ and then rearrange again.

Here's an example.

Example 1.23. We express $93/35$ as a continued fraction.

Solution. We write

$$\begin{aligned}
\frac{93}{35} &= 2 + \frac{23}{35} \\
&= 2 + \frac{1}{35/23} \\
&= 2 + \frac{1}{1 + \frac{12}{23}} \\
&= 2 + \frac{1}{1 + \frac{1}{23/12}} \\
&= 2 + \cfrac{1}{1 + \cfrac{1}{1 + \frac{11}{12}}} \\
&= 2 + \cfrac{1}{1 + \cfrac{1}{1 + \frac{1}{12/11}}} \\
&= 2 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \frac{1}{11}}}},
\end{aligned}$$

so $\frac{93}{35} = [2; 1, 1, 1, 11]$. ■

Exercise 1.24. Express $47/31$ as a continued fraction.

Compare Example 1.16 with Example 1.23: the coefficients $[2; 1, 1, 1, 11]$ match up exactly with the quotients appearing in the Euclidean algorithm. Rigorizing this is a little technical, but it is not too hard.

Proposition 1.25. Let a_0 and a_1 be coprime positive integers, and define integer sequences $q_0, q_1, \ldots, q_N$ and $a_0, a_1, a_2, \ldots, a_N$ recursively as in Proposition 1.18 by

$$a_n = q_n a_{n+1} + a_{n+2}$$

for any $n \geq 0$, where $0 < a_{n+2} < a_{n+1}$ and terminating once $a_N = 1$ so that $a_{N+1} = 0$. Then $\frac{a_0}{a_1} = [q_0; q_1, q_2, \ldots, q_N]$.

Proof. Recall that N exists by the Euclidean algorithm. We induct on N. If $N = 1$, then $a_1 = 1$ and

$$a_0 = q_0 a_1 + a_2$$

forces $a_2 = 0$ and $q_0 = a_0$. Thus, $a_0 = \frac{a_0}{a_1} = q_0 = [q_0]$.

Now take $N > 1$ (which implies $a_2 > 0$), and suppose the statement is true at $N - 1$. Then we see $a_0 = q_0 a_1 + a_2$ implies

$$\frac{a_0}{a_1} = q_0 + \frac{1}{a_1/a_2}.$$

Thus, running the Euclidean algorithm with the coprime positive integers a_1 and a_2, we find that $\frac{a_1}{a_2} = [q_1; q_2, \ldots, q_N]$ by the inductive hypothesis. It follows

$$\frac{a_0}{a_1} = q_0 + \frac{1}{[q_1; q_2, \ldots, q_N]} = [q_0; q_1, q_2, \ldots, q_N],$$

which is what we wanted. ∎

Remark 1.26. Proposition 1.25 also has the nice side effect of showing that any rational number is equal to some finite continued fraction. However, note this continued fraction is not unique: given integers $a_0, a_1, a_2, \ldots, a_n$ with $a_1, a_2, \ldots, a_n$ positive, one has

$$[a_0; a_1, a_2, \ldots, a_{n-1}, a_n] = [a_0; a_1, a_2, \ldots, a_{n-1}, a_n - 1, 1]$$

when $a_n > 1$, and otherwise

$$[a_0; a_1, a_2, \ldots, a_{n-1}, 1] = [a_0; a_1, a_2, \ldots, a_{n-1} + 1].$$

In particular, given any rational number q, we can find n of any parity such that there are integers $a_0, a_1, a_2, \ldots, a_n$ with $a_1, a_2, \ldots, a_n$ positive and $q = [a_0; a_1, a_2, \ldots, a_n]$.

The proof of Proposition 1.25 is fairly instructive: many of our arguments involving continued fractions are going to be inductive ones using identities like

$$q_0 + \frac{1}{[q_1; q_2, \ldots, q_N]} = [q_0; q_1, q_2, \ldots, q_N].$$

1.2.2 Continued Fraction Convergents

We mentioned at the outset that continued fractions provide good rational approximations for numbers. The way that this is done is by taking a long continued fraction $[a_0; a_1, a_2, \ldots]$ and "truncating" it at some point to produce the shorter (and notably finite) continued fraction $[a_0; a_1, a_2, \ldots, a_n]$. This truncation process is so important it has a name.

Definition 1.27 (convergent)**.** Given a continued fraction $[a_0; a_1, a_2, \ldots]$ and some $n \geq 0$, the truncation $[a_0; a_1, a_2, \ldots, a_n]$ is the *nth convergent*, often denoted

$$\frac{h_n}{k_n} \coloneqq [a_0; a_1, \ldots, a_n].$$

As usual, we begin with an example.

Example 1.28. We compute the continued fraction convergents of $93/35$.

Solution. In Example 1.23, we computed that $\frac{93}{35} = [2; 1, 1, 1, 11]$, so here are our convergents.

- The zeroth convergent is $[2] = 2$.
- The first convergent is $[2; 1] = 2 + \frac{1}{1} = 3$.
- The second convergent is $[2; 1, 1] = 2 + \frac{1}{1+1} = \frac{5}{2}$.
- The third convergent is $[2; 1, 1, 1]$ is

$$[2; 1, 1, 1] = 2 + \frac{1}{1 + \dfrac{1}{1+1}} = 2 + \frac{1}{3/2} = \frac{8}{3}.$$

- The fourth convergent is $[2; 1, 1, 1, 11] = \frac{93}{35}$. ∎

Exercise 1.29. Compute the continued fraction convergents of $47/31$.

The process outlined in Example 1.28 is rather annoying to execute by hand. We did not even compute $[2; 1, 1, 1, 11]$ by hand, but already $[2; 1, 1, 1]$ required some focus. In general, the problem with computing these convergents is that we are basically doing a totally new computation for every convergent.

However, there is a much faster way to compute these convergents: the "magic box" algorithm. For a sense of wonder, we will describe the algorithm first and then prove that it works second. We begin with the following grid.

		2	1	1	1	11
0	1					
1	0					

Explicitly, the 0s and 1s on the leftmost two columns will always be there in all computations, and the top row is made of our coefficients $[2; 1, 1, 1, 11]$. We now fill in the grid column-by-column, moving from left to right. For the next leftmost column, we multiply the coefficient 2 by the previous column and then add the column before that. In other words, we compute

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix} + 2 \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix},$$

so the next column in our grid is as follows.

		2	1	1	1	11
0	1	2				
1	0	1				

Indeed, $2/1$ is the zeroth convergent. We now repeat the process: multiply 1 by the previous column and then add the column before that, writing

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} + 1 \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix},$$

making our grid as follows.

		2	1	1	1	11
0	1	2	3			
1	0	1	2			

Indeed, $3/1$ is the first convergent. We now fill in the rest of the grid.

		2	1	1	1	11
0	1	2	3	5	8	93
1	0	1	1	2	3	35

And indeed, we see the remaining convergents $5/2$, $8/3$, and $93/35$ appear from our grid.

Exercise 1.30. Execute this "magic box" algorithm to compute the continued fraction convergents of $47/31$.

Exercise 1.31. Compute the following 2×2 "minors" of our grid, as follows.

$$\det \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \det \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, \quad \det \begin{bmatrix} 2 & 3 \\ 1 & 1 \end{bmatrix}, \quad \det \begin{bmatrix} 3 & 5 \\ 1 & 2 \end{bmatrix}, \quad \dots.$$

Do you see any patterns?

Proving that the magic box algorithm works is again somewhat technical. Perhaps the main difficulty is figuring out how to state the result, but the proof is still tricky. For now, we will settle for the following statement, but we will establish the refinement Corollary 1.36 shortly.

Proposition 1.32 (magic box)**.** Let $a_0, a_1, a_2, \dots$ be real numbers, where $a_1, a_2, \dots$ are positive. Define the sequences $\{h_n\}_{n=-2}^{\infty}$ and $\{k_n\}_{n=-2}^{\infty}$ of real numbers recursively by

$$\begin{bmatrix} h_{-2} & h_{-1} \\ k_{-2} & k_{-1} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix} h_{n+2} \\ k_{n+2} \end{bmatrix} = a_{n+2} \begin{bmatrix} h_{n+1} \\ k_{n+1} \end{bmatrix} + \begin{bmatrix} h_n \\ k_n \end{bmatrix}$$

for $n \geq -2$. Then

$$[a_0; a_1, \dots, a_n] = \frac{h_n}{k_n}$$

for any $n \geq 0$.

Proof. The requirement that $a_1, a_2, \dots$ be positive is entirely to avoid division by zero errors. We also take a moment to recognize that the $a_\bullet$ are being allowed to be real numbers rather than only integers. This will actually be relevant to the proof!

We induct on n. For $n = 0$, we can compute that $(h_0, k_0) = a_0(1, 0) + (0, 1) = (a_0, 1)$, so $\frac{h_0}{k_0} = a_0 = [a_0]$. For $n = 1$, we can compute that $(h_1, k_1) = a_1(a_0, 1) + (1, 0) = (a_1 a_0 + 1, a_1)$, so

$$\frac{h_1}{k_1} = \frac{a_1 a_0 + 1}{a_1} = a_0 + \frac{1}{a_1} = [a_0; a_1].$$

Now take $n \geq 2$. The trick for the inductive step is to write

$$[a_0; a_1, \dots, a_{n-2}, a_{n-1}, a_n] = a_0 + \cfrac{1}{a_1 + \cfrac{1}{\ddots + a_{n-2} + \cfrac{1}{a_{n-1} + \cfrac{1}{a_n}}}} = \left[a_0; a_1, \dots, a_{n-2}, a_{n-1} + \frac{1}{a_n}\right].$$

We may now apply the inductive hypothesis to this altered continued fraction, which is legal because $a_{n-1} + 1/a_n$ is surely a positive real number. Explicitly, define the sequence $a'_0, a'_1, \ldots, a'_{n-1}$ where $a'_m := a_m$ for $m < n-1$ and $a'_{n-1} := a_{n-1} + \frac{1}{a_n}$, and then define the sequence $\{h'_m\}_{m=-2}^{n-1}$ and $\{k'_m\}_{m=-2}^{\infty}$ as in the proposition so that

$$[a_0; a_1, \ldots, a_{n-2}, a_{n-1}, a_n] = [a'_0; a'_1, \ldots, a'_{n-1}] = \frac{h'_{n-1}}{k'_{n-1}}.$$

To compute h'_{n-1} and k'_{n-1} we acknowledge that the construction of the $a'_\bullet$ implies that $h'_m = h_m$ and $k'_m = k_m$ for $m < n-1$. So we see that

$$\begin{aligned}
\begin{bmatrix} h'_{n-1} \\ k'_{n-1} \end{bmatrix} &= a'_{n-1} \begin{bmatrix} h'_{n-2} \\ k'_{n-2} \end{bmatrix} + \begin{bmatrix} h'_{n-3} \\ k'_{n-3} \end{bmatrix} \\
&= \left(a_{n-1} + \frac{1}{a_n} \right) \begin{bmatrix} h_{n-2} \\ k_{n-2} \end{bmatrix} + \begin{bmatrix} h_{n-3} \\ k_{n-3} \end{bmatrix} \\
&= \begin{bmatrix} \left(a_{n-1} + \frac{1}{a_n} \right) h_{n-2} + h_{n-3} \\ \left(a_{n-1} + \frac{1}{a_n} \right) k_{n-2} + k_{n-3} \end{bmatrix}.
\end{aligned}$$

From here, we compute

$$\begin{aligned}
\frac{h'_{n-1}}{k'_{n-1}} &= \frac{a_{n-1}a_n h_{n-2} + h_{n-2} + a_n h_{n-3}}{a_{n-1}a_n k_{n-2} + k_{n-2} + a_n k_{n-3}} \\
&= \frac{a_n(a_{n-1}h_{n-2} + h_{n-3}) + h_{n-2}}{a_n(a_{n-1}k_{n-2} + k_{n-3}) + k_{n-2}} \\
&= \frac{a_n h_{n-1} + h_{n-2}}{a_n k_{n-1} + k_{n-2}} \\
&= \frac{h_n}{k_n},
\end{aligned}$$

which completes the proof. ■

Remark 1.33. The proof of Proposition 1.32 in fact works even if we merely assume that the $a_\bullet$ are indeterminate variables.

Example 1.34. Define the Fibonacci sequence $\{F_n\}_{n=0}^{\infty}$ by $F_0 = 0$ and $F_1 = 1$ and $F_{n+2} = F_{n+1} + F_n$ for any $n \geq 0$. Then for any $n \geq 0$,

$$\underbrace{[1; 1, \ldots, 1]}_{n+1} = \frac{F_{n+2}}{F_{n+1}}.$$

Solution. We proceed by induction on n, using Proposition 1.32. From there, we may compute that $h_0/k_0 = 1/1 = F_2/F_1$ and $h_1/k_1 = 2/1 = F_3/F_2$. For the inductive step, we note that Proposition 1.32 yields

$$h_{n+2} = h_{n+1} + h_n \qquad \text{and} \qquad k_{n+2} = k_{n+1} + k_n$$

for any $n \geq 0$, which is the recursion for the Fibonacci sequence. ■

1.2.3 More on the Magic Box Algorithm

Proposition 1.32 essentially explains why the magic box works, though perhaps there is some doubt that the fractions h_n/k_n is in reduced form. Let's show this. We begin by explaining Exercise 1.31.

Corollary 1.35. Let $a_0, a_1, a_2, \ldots$ be real numbers, where $a_1, a_2, \ldots$ are positive, and define $\{h_n\}_{n=-2}^{\infty}$ and $\{k_n\}_{n=-2}^{\infty}$ as in Proposition 1.32. Then

$$\det \begin{bmatrix} h_n & h_{n+1} \\ k_n & k_{n+1} \end{bmatrix} = (-1)^{n+1}$$

for any $n \geq -2$.

Proof. This is essentially row-reduction. We proceed by induction on n. At $n = -2$, we see that $\det \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right] = -1$. For the inductive step, suppose the statement for n, and we show $n+1$. We note

$$\begin{bmatrix} h_{n+2} \\ k_{n+2} \end{bmatrix} = a_{n+2} \begin{bmatrix} h_{n+1} \\ k_{n+1} \end{bmatrix} + \begin{bmatrix} h_n \\ k_n \end{bmatrix}$$

allows us to use column operations in order to see

$$\det \begin{bmatrix} h_{n+1} & h_{n+2} \\ k_{n+1} & k_{n+2} \end{bmatrix} = \det \begin{bmatrix} h_{n+1} & h_n \\ k_{n+1} & k_n \end{bmatrix} = -\det \begin{bmatrix} h_n & h_{n+1} \\ k_n & k_{n+1} \end{bmatrix} = -(-1)^{n+1} = (-1)^{n+2},$$

which is what we wanted. ■

Corollary 1.36. Let $a_0, a_1, a_2, \ldots$ be integers, where $a_1, a_2, \ldots$ are positive, and define $\{h_n\}_{n=-2}^{\infty}$ and $\{k_n\}_{n=-2}^{\infty}$ as in Proposition 1.32. Then, for any $n \geq 0$,

$$[a_0; a_1, \ldots, a_n] = \frac{h_n}{k_n},$$

and h_n/k_n is a fraction in reduced form with $k_n \geq 1$.

Proof. The equality follows directly from Proposition 1.32. Additionally, note that h_n and k_n are integers because they are terms of a sequence defined by integer recursion. Thus, to complete the proof, we must show that $\gcd(h_n, k_n) = 1$ and that $k_n \geq 1$ for $n \geq 0$. On one hand, we see $\gcd(h_n, k_n) = 1$ is direct from Corollary 1.35. On the other hand, $k_n \geq 1$ follows from a quick induction because $k_{-1} = 0$ and $k_0 = a_1 \geq 1$ and so $k_{n+2} = a_{n+2}k_{n+1} + k_n \geq 1$ always. ■

Corollary 1.35 has in fact suggested a faster algorithm (in terms of memory) than the Extended Euclidean algorithm. Let's see this by example.

Example 1.37. We find integers x and y such that $93x + 35y = 1$.

Solution. As in Example 1.16, we begin by writing

$$\begin{aligned} 93 &= 2 \cdot 35 + 23 \\ 35 &= 1 \cdot 23 + 12 \\ 23 &= 1 \cdot 12 + 11 \\ 12 &= 1 \cdot 11 + 1 \\ 11 &= 11 \cdot 1 + 0. \end{aligned}$$

From here, we apply the magic box algorithm Proposition 1.32 to build the following grid.

		2	1	1	1	11
0	1	2	3	5	8	93
1	0	1	1	2	3	35

Tracking Corollary 1.35 through, we see that

$$35 \cdot 8 - 93 \cdot 3 = \det \begin{bmatrix} 8 & 93 \\ 3 & 25 \end{bmatrix} = 1,$$

so $(x, y) = (-3, 8)$ works. ■

Remark 1.38. Here are a few ways to "check" the magic box algorithm.

- If using the magic box algorithm to compute convergents of the fraction p/q, then the last column of the magic box grid should yield p/q.
- The magic box algorithm has 2×2 minors controlled by Corollary 1.35, so one can compute a few of these for security.

1.2.4 Problems

Do at least 10 points worth of the following exercises.

Problem 1.2.1 (1 point). Find integer sequences $a_0, a_1, a_2, \ldots, a_m$ and $b_0, b_1, b_2, \ldots, b_n$ with $a_1, a_2, \ldots, a_m$ and $b_1, b_2, \ldots, b_n$ positive such that the sequences are distinct, but

$$[a_0; a_1, \ldots, a_m] = [b_0; b_1, \ldots, b_n].$$

Problem 1.2.2 (2 points). Compute the continued fraction convergents of $1738/1027$.

Problem 1.2.3 (3 points). Let $a_0, a_1, a_2, \ldots$ be integers, where $a_1, a_2, \ldots$ are positive, and define $\{h_n\}_{n=-2}^{\infty}$ and $\{k_n\}_{n=-2}^{\infty}$ as in Proposition 1.32. Show that

$$\left| \det \begin{bmatrix} h_n & h_{n+2} \\ k_n & k_{n+2} \end{bmatrix} \right| = |a_{n+2}|$$

for any $n \geq -2$. Additionally, predict the sign as a function on n.

Problem 1.2.4 (5 or 6 points). Let $a_0, a_1, a_2, \ldots, a_m$ and $b_0, b_1, b_2, \ldots, b_n$ be integers with $a_1, a_2, \ldots, a_m$ and $b_1, b_2, \ldots, b_n$ positive. Suppose

$$[a_0; a_1, a_2, \ldots, a_m] = [b_0; b_1, b_2, \ldots, b_n].$$

(a) For five points, suppose $m = n$. Show that $a_k = b_k$ for all $0 \leq k \leq m$.

(b) For an additional point, suppose $m < n$. Show that $m = n - 1$ and $a_k = b_k$ for $0 \leq k \leq m - 1$.

Problem 1.2.5 (5 points). Write (and submit) a function in Python which takes as input a list of integers $[a_0, a_1, a_2, \ldots]$ with $a_1, a_2, \ldots$ positive and an index n and outputs the nth convergent $[a_0; a_1, a_2, \ldots, a_n]$. You should implement the magic box algorithm.

Your test case is $[2; 1, 2, 1, 1, 4, 1, 1, 6, 1]$.

1.3 Infinite Continued Fractions

In this section, we examine continued fractions more closely. Our main task will be to show that continued fractions provide good and in fact the best rational approximations for a given irrational number. Of course, it will be a nontrivial task in order to make sense of what "best" means in this context. To set up our intuition, we will say that a fraction h/k provides a good rational approximation for a real number α if the difference

$$\left|\alpha - \frac{h}{k}\right|$$

is smaller than one might expect it to be. Of course, for any given denominator, we know that $\lfloor k\alpha \rfloor \leq k\alpha < \lfloor k\alpha \rfloor + 1$, so

$$\left|\alpha - \frac{\lfloor k\alpha \rfloor}{k}\right| \leq \frac{1}{k},$$

so a bound of $1/k$ is not too impressive. In fact, if α is irrational, we will be able to show that there are infinitely many rational numbers h/k such that

$$\left|\alpha - \frac{h}{k}\right| < \frac{1}{\sqrt{5}k^2},$$

and we will be able to show that this bound is essentially sharp.

1.3.1 Convergence of Infinite Continued Fractions

Thus far our discussion has been focused on finite continued fractions. We would now like to extend this discussion to infinite continued fractions. As in Remark 1.22, we would like to define

$$[a_0; a_1, a_2, \ldots] \stackrel{?}{:=} \lim_{n\to\infty} [a_0; a_1, a_2, \ldots, a_n],$$

but we should begin by showing that this limit in fact exists. The idea is to show that the infinite continued fraction is an infinite series, and then we can use known results on infinite series to complete the proof. As such, we begin by turning $[a_0; a_1, a_2, \ldots]$ into a series.

Lemma 1.39. Let $a_0, a_1, a_2, \ldots$ be real numbers, where $a_1, a_2, \ldots$ are positive, and let $\{h_n/k_n\}_{n=0}^{\infty}$ denote the continued fraction convergents $h_n/k_n := [a_0; a_1, a_2, \ldots, a_n]$ where $k_n \geq 1$ and $\gcd(h_n, k_n) = 1$. Then

$$\frac{h_n}{k_n} - \frac{h_{n+1}}{k_{n+1}} = \frac{(-1)^{n+1}}{k_n k_{n+1}}.$$

Thus,

$$\frac{h_n}{k_n} = \frac{h_0}{k_0} + \sum_{m=0}^{n-1} \frac{(-1)^m}{k_m k_{m+1}}.$$

Proof. Note that $\{h_n\}_{n=0}^{\infty}$ and $\{k_n\}_{n=0}^{\infty}$ are the sequences constructed in Proposition 1.32 by Corollary 1.36. As such, the first claim follows directly from Corollary 1.35. The second claim now follows from writing

$$\frac{h_n}{k_n} = \frac{h_0}{k_0} + \sum_{m=0}^{n-1} \left(\frac{h_{m+1}}{k_{m+1}} - \frac{h_m}{k_m}\right) = \frac{h_0}{k_0} + \sum_{m=0}^{n-1} \frac{(-1)^m}{k_m k_{m+1}},$$

which is what we wanted. ∎

Proposition 1.40. Let $a_0, a_1, a_2, \ldots$ be integers, where $a_1, a_2, \ldots$ are positive, and let $\{h_n/k_n\}_{n=0}^{\infty}$ denote the continued fraction convergents $h_n/k_n \coloneqq [a_0; a_1, \ldots, a_n]$ where $k_n \geq 1$ and $\gcd(h_n, k_n) = 1$. Then

$$\alpha \coloneqq \lim_{n\to\infty} [a_0; a_1, a_2, \ldots, a_n]$$

converges, and

$$\frac{1}{k_n(k_{n+1}+k_n)} < \left|\alpha - \frac{h_n}{k_n}\right| < \frac{1}{k_n k_{n+1}}$$

for each $n \geq 0$.

Proof. As usual, note that $\{h_n\}_{n=0}^{\infty}$ and $\{k_n\}_{n=0}^{\infty}$ are the sequences constructed in Proposition 1.32 by Corollary 1.36. To begin, we compute the limit as

$$\alpha = \lim_{n\to\infty} \frac{h_n}{k_n} = \frac{h_0}{k_0} + \sum_{n=0}^{\infty} \frac{(-1)^n}{k_n k_{n+1}},$$

where we have used Lemma 1.39 in the last equality. Now, the sequence $\{k_n\}_{n=0}^{\infty}$ is strictly increasing by Proposition 1.32 because $a_1, a_2, \ldots$ are all positive integers. Thus, the summation above absolute converges: an induction shows $k_n \geq n+1$, so

$$\frac{h_0}{k_0} + \sum_{n=0}^{\infty} \left|\frac{(-1)^n}{k_n k_{n+1}}\right| \leq \frac{h_0}{k_0} + \sum_{n=0}^{\infty} \frac{1}{(n+1)(n+2)} < \infty.$$

As such, the limit does in fact converge.

To compute the error term, we use the error bound for alternating series. To begin the computation, note that the above work allows us to write

$$\left|\alpha - \frac{h_n}{k_n}\right| = \left|\frac{h_0}{k_0} + \sum_{m=0}^{\infty} \frac{(-1)^m}{k_m k_{m+1}} - \frac{h_0}{k_0} - \sum_{m=0}^{n-1} \frac{(-1)^m}{k_m k_{m+1}}\right| = \left|\sum_{m=n}^{\infty} \frac{(-1)^m}{k_m k_{m+1}}\right|.$$

Because the sequence $\{k_m\}_{m=0}^{\infty}$ is strictly increasing, the terms in the sum are monotonously decreasing in magnitude to zero, so the error bound for alternating series forces $|\alpha - h_n/k_n| < 1/(k_n k_{n+1})$, which proves the upper bound for our error.

To prove the lower bound of the error, we adjust for signs and note that the sum is

$$\begin{aligned}
\left|\alpha - \frac{h_n}{k_n}\right| &= \left|\sum_{m=0}^{\infty} \frac{(-1)^m}{k_{m+n}k_{m+n+1}}\right| \\
&= \left|\sum_{m=0}^{\infty} \left(\frac{1}{k_{2m+n}k_{2m+n+1}} - \frac{1}{k_{2m+n+1}k_{2m+n+2}}\right)\right| \\
&= \left|\sum_{m=0}^{\infty} \frac{1}{k_{2m+n+1}} \cdot \frac{k_{2m+n+2} - k_{2m+n}}{k_{2m+n}k_{2m+n+2}}\right|.
\end{aligned}$$

Because $\{k_n\}_{n=0}^{\infty}$ is a strictly increasing sequence, all the terms of the sum are positive, so we may remove the absolute signs to see

$$\left|\alpha - \frac{h_n}{k_n}\right| > \frac{1}{k_{n+1}} \cdot \frac{k_{n+2} - k_n}{k_n k_{n+2}}.$$

Thus, to prove the desired lower bound, it suffices to show $k_{n+1}k_{n+2} \leq (k_{n+1} + k_n)(k_{n+2} - k_n)$. This rearranges to $k_n(k_n + k_{n+1}) \leq k_n k_{n+2}$, or $k_n + k_{n+1} \leq k_{n+2}$, which is true by Proposition 1.32. ∎

Remark 1.41. Proposition 1.40 tells us that h_n/k_n will be a "better" rational approximation for α when k_{n+1} is particularly large. For example, $\pi = [3; 7, 15, 1, 292, 1, 1, 1]$, so we would guess that

$$[3; 7, 15, 1] = \frac{355}{113} = 3.14159292035\ldots$$

is a particularly good rational approximation of π, and indeed it is. Notably, $[3; 7] = 22/7$ is also a remarkable rational approximation.

As such, we may make the following definition.

Definition 1.42 (infinite continued fraction)**.** Let $a_0, a_1, a_2, \ldots$ be integers, where $a_1, a_2, \ldots$ are positive. Then we define the *infinite continued fraction*

$$[a_0; a_1, a_2, \ldots] \coloneqq \lim_{n\to\infty} [a_0; a_1, a_2, \ldots, a_n].$$

Example 1.43. We have

$$\varphi \coloneqq \frac{1+\sqrt{5}}{2} = [1; 1, 1, \ldots].$$

Solution. By Proposition 1.40, we know that $[1; 1, 1, \ldots]$ converges to some real number α. Further,

$$\alpha = 1 + \cfrac{1}{1 + \cfrac{1}{1 + \ddots}} = 1 + \frac{1}{\alpha},$$

which rearranges to $\alpha^2 - \alpha - 1 = 0$, so

$$\alpha \in \left\{ \frac{1 \pm \sqrt{5}}{2} \right\}.$$

However, we claim that $\alpha > 0$. With the tools we have, this is somewhat annoying to show, but we remark that Lemma 1.61 makes this relatively easy. Anyway, let $\{h_n/k_n\}_{n=0}^{\infty}$ denote the continued fraction convergents. Proposition 1.32 implies that $h_0/k_0 = 1/1$ and $h_1/k_1 = 2/1$, so

$$|\alpha - 1| = \left|\alpha - \frac{h_0}{k_0}\right| < \frac{1}{k_0 k_1} = 1,$$

so $\alpha > 0$. Thus, $\alpha = \varphi$. ∎

Exercise 1.44. Compute $[2; 2, 2, \ldots]$.

The above examples have the amusing feature that $[a_0; a_1, a_2, \ldots]$ is irrational. This is not a coincidence. The following result is perhaps our first "Diophantine approximation" result.

Proposition 1.45. Let α be a real number, and let $C > 0$ and $\varepsilon > 0$. Then α is irrational if there is a sequence of rational numbers $\{h_n/k_n\}_{n=0}^{\infty}$ such that

$$\left|\alpha - \frac{h_n}{k_n}\right| < \frac{C}{k_n^{1+\varepsilon}}$$

for each $n \geq 0$.

Proof. We show the contrapositive. Suppose that $\alpha = p/q$ is rational with $q \geq 1$ and $\gcd(p,q) = 1$, and we show that there are only finitely many rational numbers h/k such that $|\alpha - h/k| < C/k^{1+\varepsilon}$; we may assume that $k \geq 1$ and that $\gcd(h,k) = 1$ in our fractions h/k. Now, for any given k, we note that our inequality rearranges to

$$|h - k\alpha| < \frac{C}{k^\varepsilon},$$

so there are only finitely many integers h in our interval. Thus, it suffices to upper-bound k. Well, plugging in $\alpha = p/q$ and clearing fractions reveals that we want

$$|qh - pk| < \frac{Cq}{k^\varepsilon}.$$

Now, we claim that $k \leq \max\left\{(Cq)^{1/\varepsilon}, q\right\}$, which completes the proof. Well, suppose that $k^\varepsilon > Cq$, and we will show $k = q$. Indeed, $qh - pk$ is an integer with magnitude less than 1, so it follows that $qh - pk = 0$, so in fact

$$qh = pk.$$

By the uniqueness of our representation of rational numbers, it follows that $k = q$. Explicitly, $q \mid pk$, but $\gcd(q,p) = 1$, so $q \mid k$. A symmetric argument shows $k \mid q$, so $k, q \geq 1$ establishes $k = q$. ∎

Remark 1.46. Proposition 1.45 is fairly surprising result! Approximately speaking, it says that having "too many" good rational approximations of a given real number actually forces the real number to be irrational! We will prove a converse shortly in Corollary 1.53.

Remark 1.47. Here is a way to intuit Proposition 1.45: there is a sense in which rational numbers cannot be "too close to each other" simply because

$$\left|\frac{a}{b} - \frac{c}{d}\right| \geq \frac{1}{|bd|}.$$

Thus, we should not be able to use rational numbers to provide good rational approximations of rational numbers.

Corollary 1.48. Let $a_0, a_1, a_2, \ldots$ be integers, where $a_1, a_2, \ldots$ are positive. Then $[a_0; a_1, a_2, \ldots]$ is irrational.

Proof. Let $\{h_n/k_n\}_{n=0}^\infty$ denote the continued fraction convergents $h_n/k_n := [a_0; a_1, \ldots, a_n]$ where $k_n \geq 1$ and $\gcd(h_n, k_n) = 1$. Then Proposition 1.40 establishes that

$$\left|[a_0; a_1, a_2, \ldots] - \frac{h_n}{k_n}\right| < \frac{1}{k_n k_{n+1}} < \frac{1}{k_n^2}$$

for each $n \geq 0$, where the last inequality follows because $\{k_n\}_{n=0}^\infty$ is strictly increasing. Proposition 1.45 completes the proof. ∎

1.3.2 Building Infinite Continued Fractions

Given an irrational real number $\alpha \in \mathbb{R} \setminus \mathbb{Q}$, we would like to construct a sequence of integers $a_0, a_1, a_2, \ldots$ with $a_1, a_2, \ldots$ positive and $\alpha = [a_0; a_1, a_2, \ldots]$. We did this by hand for φ in Example 1.43, but this is not a general algorithm.

Let's describe what the algorithm should be. Suppose we could write $\alpha = [a_0; a_1, a_2, \ldots]$. Then

$$\alpha = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots}}$$

forces $a_0 = \lfloor \alpha \rfloor$. From here, define $\alpha_1 \coloneqq (\alpha - a_0)^{-1}$, and we see

$$\alpha_1 = a_1 + \cfrac{1}{a_2 + \ddots}.$$

Then we can see that we must have $a_1 = \lfloor \alpha_1 \rfloor$, and we go on to define $\alpha_2 = (\alpha_1 - a_1)^{-1}$ and continue the process. This suggests the following result.

Proposition 1.49. Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ be an irrational number. Define the sequence of real numbers $\{\alpha_n\}_{n=0}^{\infty}$ and integers $\{a_n\}_{n=0}^{\infty}$ by $\alpha_0 \coloneqq \alpha$ and

$$a_n \coloneqq \lfloor \alpha_n \rfloor \qquad \text{and} \qquad \alpha_{n+1} \coloneqq \frac{1}{\alpha_n - a_n}$$

Then $a_0, a_1, a_2, \ldots$ are integers, and $a_1, a_2, \ldots$ are positive, and $\alpha = [a_0; a_1, a_2, \ldots]$.

Proof. Quickly, we note that there are no division by zero problems: by construction, the $a_\bullet$ are all integers, and the recursion implies that α_{n+1} is irrational if and only if α_n is irrational, so induction implies that all the $\alpha_\bullet$ are irrational. Next up, we note that $a_n < \alpha_n < a_n + 1$ for each $n \geq 0$ (recall α_n is irrational for each n), so $0 < \alpha_n - a_n < 1$ for each $n \geq 0$, so $a_{n+1} \geq 1$ for each $n \geq 0$, so $a_1, a_2, \ldots$ are in fact positive integers.

It remains to show $\alpha = [a_0; a_1, a_2, \ldots]$. This is somewhat technical. The main claim is that

$$\alpha \overset{?}{=} [a_0; a_1, \ldots, a_n, \alpha_{n+1}]$$

for each $n \geq 0$. We show this by induction. For $n = -1$, there is nothing to say because $\alpha = \alpha_0$. For the induction, we write

$$\begin{aligned}
\alpha &= [a_0; a_1, \ldots, a_n, \alpha_{n+1}] \\
&= [a_0; a_1, \ldots, a_n, \lfloor \alpha_{n+1} \rfloor + \{\alpha_{n+1}\}] \\
&= \left[a_0; a_1, \ldots, a_n, a_{n+1} + \frac{1}{\alpha_{n+2}}\right] \\
&= [a_0; a_1, \ldots, a_n, a_{n+1}, \alpha_{n+2}],
\end{aligned}$$

which completes the induction.

We now finish the proof that $\alpha = [a_0; a_1, a_2, \ldots]$. For each $n \geq 0$, set $h_n/k_n \coloneqq [a_0; a_1, \ldots, a_n]$ and $h'_{n+1}/k'_{n+1} \coloneqq [a_0; a_1, a_2, \ldots, a_n, \alpha_{n+1}]$ as constructed in Proposition 1.32. Then applying Lemma 1.39 implies

$$\begin{aligned}
\alpha - [a_0; a_1, a_2, \ldots, a_n] &= [a_0; a_1, \ldots, a_n, \alpha_{n+1}] - [a_0; a_1, a_2, \ldots, a_n] \\
&= \frac{h_0}{k_0} + \sum_{m=0}^{n-1} \frac{(-1)^m}{k_m k_{m+1}} - \frac{h_0}{k_0} - \sum_{m=0}^{n-1} \frac{(-1)^m}{k_m k_{m+1}} - \frac{(-1)^n}{k_n k'_{n+1}} \\
&= \frac{(-1)^n}{k_n k'_{n+1}}.
\end{aligned}$$

Thus,

$$|\alpha - [a_0; a_1, a_2, \ldots, a_n]| \leq \frac{1}{k_n^2},$$

where we have used the fact that $k'_{n+1} = \alpha_{n+1} k_n + k_{n-1} \geq k_n$. Sending $n \to \infty$ makes $k_n \to \infty$, so we conclude $[a_0; a_1, \ldots, a_n] \to \alpha$ as $n \to \infty$. ∎

Exercise 1.50. Use Proposition 1.49 (and Sage) to compute the first 10 continued fraction coefficients of π.

Remark 1.51. In contrast to Remark 1.26, the continued fraction attached to irrational $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ is unique. The proof is approximately along the lines as the argument at the start of the subsection. Namely, suppose we have integers $a_0, a_1, a_2, \ldots$ and $b_0, b_1, b_2, \ldots$ with $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$ positive, and suppose

$$[a_0; a_1, a_2, \ldots] = [b_0; b_1, b_2, \ldots].$$

We want to show $a_n = b_n$ for all n. Because $[a_0; a_1, a_2, \ldots] = a_0 + [a_1; a_2, \ldots]^{-1}$, it suffices by induction to show that $a_0 = b_0$. Well, $a_1, b_1 \geq 1$ implies $[a_1; a_2, \ldots], [b_1, b_2, \ldots] > 1$, so

$$a_0 = \left\lfloor a_0 + \frac{1}{[a_1; a_2, \ldots]} \right\rfloor = \lfloor [a_0; a_1, a_2, \ldots] \rfloor = \lfloor [b_0; b_1, b_2, \ldots] \rfloor = b_0.$$

Proposition 1.49 allows us to make the following terminology.

Definition 1.52 (convergent)**.** Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ be an irrational number. By Proposition 1.49, we may find integers $a_0, a_1, a_2, \ldots$ where $a_1, a_2, \ldots$ are positive and $\alpha = [a_0; a_1, a_2, \ldots]$. Then the *$n$th continued fraction convergent of* α is $[a_0; a_1, a_2, \ldots, a_n]$.

Corollary 1.53. Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ be an irrational number. Then there is a sequence of rational numbers $\{h_n/k_n\}_{n=0}^{\infty}$ such that

$$\left| \alpha - \frac{h_n}{k_n} \right| < \frac{1}{k_n^2}$$

for each $n \geq 0$.

Proof. We use continued fraction convergents. Let $\{h_n/k_n\}_{n=0}^{\infty}$ be the sequence of continued fraction convergents of α. Then Proposition 1.40 implies

$$\left| \alpha - \frac{h_n}{k_n} \right| < \frac{1}{k_n k_{n+1}}.$$

Because $k_{n+1} > k_n$ by the recursion, the conclusion follows. ∎

1.3.3 Quadratic Irrationals

As an intermission, we take a moment to compute the continued fraction of $\sqrt{d}$ where d is a non-square positive integer. Let's start with an example.

Example 1.54. We show that $\sqrt{3} = [1; \overline{1, 2}]$, where the over line indicates periodicity.

Solution. It is possible to solve this by computing $[\overline{1, 2}]$ first as in Example 1.43, but we will take a more direct

approach following Proposition 1.49. Indeed, following the algorithm, we compute

$$\begin{aligned}\sqrt{3} &= 1 + \left(-1 + \sqrt{3}\right) \\ &= 1 + \frac{1}{\frac{1+\sqrt{3}}{2}} \\ &= 1 + \frac{1}{1 + \frac{-1+\sqrt{3}}{2}} \\ &= 1 + \frac{1}{1 + \frac{1}{1+\sqrt{3}}} \\ &= 1 + \frac{1}{1 + \frac{1}{2 + (-1+\sqrt{3})}}.\end{aligned}$$

At this point, we have seen that

$$\frac{1+\sqrt{3}}{2} = 1 + \frac{1}{2 + \frac{1}{\frac{1+\sqrt{3}}{2}}},$$

so $\overline{[1,2]} = \frac{1+\sqrt{3}}{2}$ follows, from which the desired result follows by noting $\sqrt{3} = 1 + 1/\left(\frac{1+\sqrt{3}}{2}\right)$ as computed above. ■

As in the above example, it will turn out that the terms α_n from Proposition 1.49 will all take the form

$$\frac{r_n + \sqrt{d}}{s_n}$$

for some positive integers r_n, s_n. We are now ready to state our result.

Proposition 1.55. Fix a non-square positive integer d. Define $a_0 := \left\lfloor \sqrt{d} \right\rfloor$ and $r_0 := 0$ and $s_0 := 1$ and $\alpha_0 := \sqrt{d}$ and

$$a_n := \left\lfloor \frac{r_n + a_0}{s_n} \right\rfloor, \qquad r_{n+1} := a_n s_n - r_n, \qquad \text{and} \qquad s_{n+1} := \frac{d - r_{n+1}^2}{s_n}$$

for each $n \geq 0$. Then these are sequences of integers with $0 \leq r_n < \sqrt{d}$ and $1 \leq s_n < 2\sqrt{d}$, and $\sqrt{d} = [a_0; a_1, a_2, \ldots]$.

Proof. We forget about the sequences defined in the statement of the proposition, and we will redefine such sequences a different way in the argument which follows. Let $\{a_n\}_{n=0}^{\infty}$ and $\{\alpha_n\}_{n=0}^{\infty}$ be as in Proposition 1.49. The main point is to compute the $\alpha_\bullet$s. We proceed in steps.

1. We define our sequences. The proof of Proposition 1.49 actually shows that $\sqrt{d} = [a_0; a_1, \ldots, a_n; \alpha_{n+1}]$ for any $n \geq 0$, so Proposition 1.32 shows that any $n \geq 0$ has

$$\sqrt{d} = \frac{h_{n-1}\alpha_n + h_{n-2}}{k_{n-1}\alpha_n + k_{n-2}},$$

where $\{h_n/k_n\}_{n=-2}^{\infty}$ are the continued fraction convergents of $\sqrt{d}$, and we have taken $(h_{-2}, k_{-2}) = (0,1)$ and $(h_{-1}, k_{-1}) = (1,0)$ formally. We can use the above equation to solve α_n as

$$\alpha_n = -\frac{h_{n-2} - k_{n-2}\sqrt{d}}{h_{n-1} - k_{n-1}\sqrt{d}} = -\frac{(h_{n-1}h_{n-2} - dk_{n-1}k_{n-2}) + (-1)^{n-1}\sqrt{d}}{h_{n-1}^2 - dk_{n-1}^2},$$

where we have used Corollary 1.35 in the last equality. As such, we define the integer sequences

$$r_n \coloneqq (-1)^n(h_{n-2}h_{n-1} - dk_{n-1}k_n)$$
$$s_n \coloneqq (-1)^n\left(h_{n-1}^2 - dk_{n-1}^2\right)$$

so that

$$\alpha_n = \frac{r_n + \sqrt{d}}{s_n}.$$

The reason we have chosen to define $r_\bullet$ and $s_\bullet$ this way is that we know that they are integer sequences "for free."

2. We show the recursions for $r_\bullet$ and $s_\bullet$. To begin, $\alpha_0 = \sqrt{d}$ forces $(r_0, s_0) = (0, 1)$. Further, the remaining recursion comes from the recursion of Proposition 1.49: write

$$\frac{r_{n+1} + \sqrt{d}}{s_{n+1}} = \alpha_{n+1} = \frac{1}{\alpha_n - a_n} = \frac{1}{\frac{r_n+\sqrt{d}}{s_n} - a_n} = \frac{s_n}{-(a_ns_n - r_n) + \sqrt{d}} = \frac{(a_ns_n - r_n) + \sqrt{d}}{\frac{d-(a_ns_n-r_n)^2}{s_n}}.$$

Thus, comparing coefficients, we see $r_{n+1} = a_ns_n - r_n$ and $s_{n+1} = \frac{1}{s_n}\left(d - r_{n+1}^2\right)$, as needed.

3. We show the inequalities on $r_\bullet$ and $s_\bullet$. Quickly, note that $h_{n-1}^2 - dk_{n-1}^2 > 0$ if and only if $h_{n-1}/k_{n-1} > \sqrt{d}$ if and only if $n-1$ is odd by Lemma 1.61, which is equivalent to n being even, meaning that $s_n > 0$ always. The recursion on $s_\bullet$ then forces $r_{n+1} < \sqrt{d}$ always, and combined with $r_0 = 0$, we see that $r_n < \sqrt{d}$ always. For the other inequalities, we show

$$0 < \frac{\sqrt{d} - r_n}{s_n} < 1$$

for $n \geq 1$. For $n = 1$, we see $r_1 = a_0$ and $s_1 = d - a_0^2$, so the given quantity is $1/\left(\sqrt{d} + a_0\right) < 1$. As for the inductive step, by replacing $\sqrt{d}$ with $-\sqrt{d}$ in the computation of the previous step, we see

$$\frac{r_{n+1} - \sqrt{d}}{s_{n+1}} = \frac{1}{\frac{r_n-\sqrt{d}}{s_n} - a_n},$$

so

$$\frac{\sqrt{d} - r_{n+1}}{s_{n+1}} = \frac{1}{\frac{\sqrt{d}-r_n}{s_n} + a_n}$$

is positive and less than 1 because $a_n \geq 1$ and the inductive hypothesis.

Now, $\alpha_n = \frac{r_n+\sqrt{d}}{s_n} > 1$ by the proof of Proposition 1.49, so adding and subtracting yields $\frac{2r_n}{s_n} > 0$ and $\frac{2\sqrt{d}}{s_n} > 1$, so $r_n > 0$ and $s_n < 2\sqrt{d}$ for $n > 0$, thus completing the step.

4. We show the recursion on $a_\bullet$. This merely requires writing

$$a_n = \lfloor \alpha_n \rfloor = \left\lfloor \frac{r_n + \sqrt{d}}{s_n} \right\rfloor = \left\lfloor \frac{r_n + a_0}{s_n} \right\rfloor,$$

so we are done. ■

Corollary 1.56. Fix a non-square positive integer d, and write $\sqrt{d} = [a_0; a_1, a_2, \ldots]$. Then there exists an integer $N \leq 2d$ and positive integer $p \leq 2d$ such that $a_{n+p} = a_n$ for each $n \geq N$.

Proof. We use Proposition 1.55. Among the d pairs $(r_0, s_0), (r_1, s_1), \dots, (r_{2d}, s_{2d})$, we must repeat some ordered pair of integers; say $(r_N, s_N) = (r_{N+p}, s_{N+p})$ for some integer $N \le 2d$ and positive integer $p < 2d$. Then the recursion

$$(r_{n+1}, s_{n+1}) = \left(\left\lfloor \frac{r_n + a_0}{s_n} \right\rfloor s_n - r_n, \frac{d - r_{n+1}^2}{s_n} \right)$$

forces $(r_n, s_n) = (r_{n+p}, s_{n+p})$ for each $n \ge N$, so $a_{n+p} = \left\lfloor \frac{r_{n+p}+a_0}{s_{n+p}} \right\rfloor = \left\lfloor \frac{r_n+a_0}{s_n} \right\rfloor = a_n$ follows. ■

Remark 1.57. Examining the above proofs, we actually see that

$$r_n := (-1)^n (h_{n-2}h_{n-1} - dk_{n-1}k_n)$$
$$s_n := (-1)^n \left(h_{n-1}^2 - dk_{n-1}^2\right)$$

enjoys the same periodicity as Corollary 1.56.

1.3.4 Convergents Are Good Rational Approximations

As before, let $a_0, a_1, a_2, \dots$ be integers, where $a_1, a_2, \dots$ are positive, and let $\{h_n/k_n\}_{n=0}^{\infty}$ denote the continued fraction convergents $h_n/k_n := [a_0; a_1, \dots, a_n]$ where $k_n \ge 1$ and $\gcd(h_n, k_n) = 1$. Proposition 1.40 immediately implies that

$$\left| \alpha - \frac{h_n}{k_n} \right| \le \frac{1}{k_n^2},$$

but we can improve this result somewhat. The goal of the present section is to show that there are infinitely many n for which

$$\left| \alpha - \frac{h_n}{k_n} \right| \le \frac{1}{\sqrt{5}k_n^2},$$

and the following example explains that the constant $\sqrt{5}$ is the best possible.

Example 1.58. Let $\varphi = \frac{1+\sqrt{5}}{2} = [1; 1, 1, \dots]$ as in Example 1.43. By Example 1.34, the nth continued fraction convergent is F_{n+2}/F_{n+1}. For any $c > \sqrt{5}$, we have

$$\left| \varphi - \frac{F_{n+2}}{F_{n+1}} \right| < \frac{1}{cF_{n+1}^2}$$

for only finitely many n.

Solution. Set $\overline{\varphi} := \frac{1-\sqrt{5}}{2}$, which is the negative solution of $x^2 = x + 1$; note $\varphi + \overline{\varphi} = 1$ and $\varphi\overline{\varphi} = -1$. An induction on n proves Binet's formula

$$F_n \stackrel{?}{=} \frac{\varphi^n - \overline{\varphi}^n}{\sqrt{5}}.$$

Indeed, the above equality holds at $n = 0$ and $n = 1$ by a direct computation, and taking a linear combination of the relations $\varphi^{n+2} = \varphi^{n+1} + \varphi^n$ and $\overline{\varphi}^{n+2} = \overline{\varphi}^{n+1} + \overline{\varphi}^n$ proves the inductive step.

We now carefully study the error. For any $n \ge 0$, we see

$$\begin{aligned} 5\left(\varphi F_{n+1}^2 - F_{n+2}F_{n+1}\right) &= \varphi\left(\varphi^{n+1} - \overline{\varphi}^{n+1}\right)^2 - \left(\varphi^{n+2} - \overline{\varphi}^{n+2}\right)\left(\varphi^{n+1} - \overline{\varphi}^{n+1}\right) \\ &= \varphi\left(\varphi^{2n+2} + \overline{\varphi}^{2n+2} - 2(\varphi\overline{\varphi})^{n+1}\right) - \left(\varphi^{2n+3} + \overline{\varphi}^{2n+3} - (\varphi\overline{\varphi})^{n+1}(\varphi + \overline{\varphi})\right) \\ &= (-1)^n(2\varphi - 1) + \overline{\varphi}^{2n+2}(\varphi - \overline{\varphi}) \\ &= (-1)^n\sqrt{5} + \overline{\varphi}^{2n+2}\sqrt{5}. \end{aligned}$$

Thus,

$$cF_{n+1}^2\left|\varphi - \frac{F_{n+2}}{F_{n+1}}\right| = \frac{c}{\sqrt{5}}\left|(-1)^n + \overline{\varphi}^{2n+2}\right|. \tag{1.1}$$

As $n \to \infty$, we see $\overline{\varphi}^{2n+2} \to 0$, so the error above approaches $c/\sqrt{5} > 1$. Thus, only finitely many n have the above quantity less than 1, which is what we wanted. ■

Remark 1.59. Carefully tracking through Example 1.58 tells us that

$$\left|\varphi - \frac{F_{n+2}}{F_{n+1}}\right| < \frac{1}{\sqrt{5}F_{n+1}^2}$$

exactly for the even n. Indeed, this follows from (1.1) upon noting $-\overline{\varphi}^{2n+2} < 0$. Compare this result with the statement and proof of Theorem 1.63.

Exercise 1.60. Set $\alpha := \sqrt{2}$, and let $\{h_n/k_n\}_{n=0}^{\infty}$ be the continued fraction convergents of α. Find the largest real number $c > 0$ for which there exist infinitely many integers $n \geq 0$ such that

$$\left|\alpha - \frac{h_n}{k_n}\right| < \frac{1}{ck_n^2}.$$

As should be somewhat evident by the $\sqrt{5}$ in our bounds and in the above proof, the arguments here are going to be somewhat ad-hoc. The following result starts us off.

Lemma 1.61. Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ be irrational, and let $\{h_n/k_n\}_{n=0}^{\infty}$ be the sequence of continued fraction convergents of α. For any $n \geq 0$, we have

$$\frac{h_{2n}}{k_{2n}} < \frac{h_{2n+2}}{k_{2n+2}} < \frac{h_{2n+3}}{k_{2n+3}} < \frac{h_{2n+1}}{k_{2n+1}}.$$

Proof. Applying Lemma 1.39, we are trying to show

$$\frac{h_{2n}}{k_{2n}} \overset{?}{<} \frac{h_{2n}}{k_{2n}} + \frac{1}{k_{2n}k_{2n+1}} - \frac{1}{k_{2n+1}k_{2n+2}} \overset{?}{<} \frac{h_{2n}}{k_{2n}} + \frac{1}{k_{2n}k_{2n+1}} - \frac{1}{k_{2n+1}k_{2n+2}} + \frac{1}{k_{2n+2}k_{2n+3}} \overset{?}{<} \frac{h_{2n}}{k_{2n}} + \frac{1}{k_{2n}k_{2n+1}}.$$

Simplifying, we want to show

$$0 \overset{?}{<} \frac{1}{k_{2n}k_{2n+1}} - \frac{1}{k_{2n+1}k_{2n+2}} \overset{?}{<} \frac{1}{k_{2n}k_{2n+1}} - \frac{1}{k_{2n+1}k_{2n+2}} + \frac{1}{k_{2n+2}k_{2n+3}} \overset{?}{<} \frac{1}{k_{2n}k_{2n+1}}.$$

The leftmost inequality is equivalent to $k_{2n} < k_{2n+2}$, which is true. The middle inequality is equivalent to $0 < 1/(k_{2n+2}k_{2n+3})$, which is true. Lastly, the rightmost inequality is equivalent to $k_{2n+1} < k_{2n+3}$, which is true. ■

Proposition 1.62. Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ be irrational, and let $\{h_n/k_n\}_{n=0}^{\infty}$ be the sequence of continued fraction convergents of α. For any $m \geq 0$, there exists $n \in \{2m, 2m+1\}$ such that

$$\left|\alpha - \frac{h_n}{k_n}\right| < \frac{1}{2k_n^2}.$$

Proof. The point is that one of h_{2m}/k_{2m} or h_{2m+1}/k_{2m+1} is going to be "closer" to α. By Lemma 1.61, we see that $\{h_{2m}/k_{2m}\}_{m=0}^{\infty}$ is a strictly ascending sequence of rational numbers, which converges to α by definition of α. Analogously, $\{h_{2m+1}/k_{2m+1}\}_{m=0}^{\infty}$ is a strictly descending sequence of rational numbers which also converges to α. Thus,

$$\frac{h_{2m}}{k_{2m}} < \alpha < \frac{h_{2m+1}}{k_{2m+1}}.$$

By Lemma 1.39, the length of this interval is $1/(k_{2m}k_{2m+1})$.

Now, suppose for contradiction that

$$\left|\alpha - \frac{h_n}{k_n}\right| \geq \frac{1}{2k_n^2}$$

for $n \in \{2m, 2m+1\}$. Then we must have

$$\frac{h_{2m}}{k_{2m}} + \frac{1}{2k_{2m}^2} \leq \alpha \leq \frac{h_{2m+1}}{k_{2m+1}} - \frac{1}{2k_{2m+1}^2}.$$

This rearranges to

$$\frac{1}{2k_{2m}^2} + \frac{1}{2k_{2m+1}^2} \leq \frac{1}{k_{2m}k_{2m+1}}$$

by Lemma 1.39, but this is equivalent to $(k_{2m} - k_{2m+1})^2 \leq 0$, or $k_{2m} = k_{2m+1}$. This is a contradiction because the sequence $\{k_n\}_{n=0}^{\infty}$ is strictly increasing. ■

With a little more care in the last half of the argument, we can achieve the desired result.

Theorem 1.63 (Hurwitz)**.** Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ be irrational, and let $\{h_n/k_n\}_{n=0}^{\infty}$ be the sequence of continued fraction convergents of α. For any $m \geq 0$, there exists $n \in \{3m, 3m+1, 3m+2\}$ such that

$$\left|\alpha - \frac{h_n}{k_n}\right| < \frac{1}{\sqrt{5}k_n^2}.$$

Proof. The proof is along the same lines as Proposition 1.62. Without loss of generality, we work with even m in order to make our inequalities better-behaved; the argument for odd m is analogous but requires reversing a few inequalities. Anyway, if m is even, Lemma 1.61 implies

$$\frac{h_{3m}}{k_{3m}} < \frac{h_{3m+2}}{k_{3m+2}} < \alpha < \frac{h_{3m+1}}{k_{3m+1}}.$$

(The location of α adjusts in the case where m is odd.) Now, suppose for the sake of contradiction that

$$\left|\alpha - \frac{h_n}{k_n}\right| \geq \frac{1}{\sqrt{5}k_n^2}$$

for each $n \in \{3m, 3m+1, 3m+2\}$. Removing the absolute values, we receive the inequalities

$$\frac{h_{3m}}{k_{3m}} + \frac{1}{\sqrt{5}k_{3m}^2} \leq \alpha, \quad \alpha \leq \frac{h_{3m+1}}{k_{3m+1}} - \frac{1}{\sqrt{5}k_{3m+1}^2}, \quad \text{and} \quad \frac{h_{3m+2}}{k_{3m+2}} + \frac{1}{\sqrt{5}k_{3m+2}^2} \leq \alpha,$$

which imply

$$\frac{h_{3m}}{k_{3m}} + \frac{1}{\sqrt{5}k_{3m}^2} \leq \frac{h_{3m+1}}{k_{3m+1}} - \frac{1}{\sqrt{5}k_{3m+1}^2}, \quad \text{and} \quad \frac{h_{3m+2}}{k_{3m+2}} + \frac{1}{\sqrt{5}k_{3m+2}^2} \leq \frac{h_{3m+1}}{k_{3m+1}} - \frac{1}{\sqrt{5}k_{3m+1}^2}.$$

By Lemma 1.39, these rearrange into

$$\frac{1}{k_{3m}^2} + \frac{1}{k_{3m+1}^2} \leq \frac{\sqrt{5}}{k_{3m}k_{3m+1}}, \quad \text{and} \quad \frac{1}{k_{3m+1}^2} + \frac{1}{k_{3m+2}^2} \leq \frac{\sqrt{5}}{k_{3m+1}k_{3m+2}}.$$

By Proposition 1.32, we see that $k_{3m} + ak_{3m+1} = k_{3m+2}$ for some integer a, so our inequalities read

$$\frac{1}{k_{3m}^2} + \frac{1}{k_{3m+1}^2} \le \frac{\sqrt{5}}{k_{3m}k_{3m+1}}, \quad \text{and} \quad \frac{1}{k_{3m+1}^2} + \frac{1}{(k_{3m} + ak_{3m+1})^2} \le \frac{\sqrt{5}}{k_{3m+1}(k_{3m} + ak_{3m+1})}.$$

Now, we set $q \coloneqq k_{3m}/k_{3m+1}$ to homogenize the inequalities. This gives

$$q^2 + 1 \le \sqrt{5}q, \quad \text{and} \quad (q+a)^2 + 1 \le \sqrt{5}(q+a).$$

In other words, we are asking for $\{q, q+a\} \subseteq \{x \in \mathbb{R} : x^2 + 1 \le \sqrt{5}x\}$. To solve for q, we note $x^2 - \sqrt{5}x + 1 = 0$ exactly when $x = \frac{\sqrt{5}\pm 1}{2}$, so $\{x \in \mathbb{R} : x^2 + 1 \le \sqrt{5}x\}$ is the closed interval from $\frac{\sqrt{5}-1}{2}$ up to $\frac{\sqrt{5}+1}{2}$. Thus, we must have $q = \frac{\sqrt{5}-1}{2}$ and $a = 1$, which is a contradiction because q is rational while $\frac{\sqrt{5}-1}{2}$ is irrational! ∎

1.3.5 Convergents Are Best Rational Approximations

Now that we are somewhat acquainted with what it means to be a "good" rational approximation, we are ready to state and prove our main result on continued fractions. It is a converse to Proposition 1.62. Our exposition in this subsection roughly follows [HW75, Theorem 184].

Theorem 1.64. Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ be irrational, and let $\{h_n/k_n\}_{n=0}^{\infty}$ be the sequence of continued fraction convergents of α. Given a rational number h/k with $\gcd(h,k) = 1$ and $k \ge 1$, if

$$\left|\alpha - \frac{h}{k}\right| < \frac{1}{2k^2},$$

then $(h,k) = (h_n, k_n)$ for some n.

Approximately speaking, Theorem 1.64 tells us that the best rational approximations of a real number are all continued fraction convergents.

Proof of Theorem 1.64. We use Remark 1.26 to write

$$\frac{h}{k} = [a_0; a_1, a_2, \dots, a_n]$$

with a_n with parity chosen so that n is even if and only if $\alpha > h/k$. (This is what we expect from Lemma 1.61.) Then let $\{p_m/q_m\}_{m=0}^{n}$ be the continued fraction convergents; for example, $(h,k) = (p_n, q_n)$.

The main idea is to show that the continued fraction expansion of α begins $[a_0; a_1, a_2, \dots, a_n, \dots]$. To realize this, we must continue the continued fraction. Well, we know that we can certainly find some $\beta \in \mathbb{R}$ such that

$$\alpha = \frac{p_n\beta + p_{n-1}}{q_n\beta + q_{n-1}}$$

by rearranging. (Explicitly, we need to know that $\alpha k - h \neq 0$ to set $\beta \coloneqq (h' - \alpha k')/(\alpha k - h)$, which is true because α is irrational.) The main claim is that $\beta > 1$. Well, comparing with our error, we see

$$\alpha - \frac{h}{k} = \frac{p_n\beta + p_{n-1}}{q_n\beta + q_{n-1}} - \frac{p_n}{q_n} = \frac{p_{n-1}q_n - p_nq_{n-1}}{(q_n\beta + q_{n-1})q_n} = \frac{(-1)^n}{(q_n\beta + q_{n-1})q_n},$$

where we applied Corollary 1.35 in the last equality. We arranged the parity n so that the left-hand side is positive if and only if $(-1)^n = 1$, so we may now write

$$1 > 2p_n^2\left|\alpha - \frac{p_n}{q_n}\right| = \frac{2p_n}{p_n\beta + p_{n-1}},$$

so $\beta > 2 - p_{n-1}/p_n$, which is bigger than 1 because $p_{n-1} < p_n$.

We now convert $\beta > 1$ into the result. Well, Proposition 1.49 allows us to write

$$\beta = [a_{n+1}; a_{n+2}, a_{n+3}, \dots]$$

for integers $a_{n+1}, a_{n+2}, a_{n+3}, \ldots$ with $a_{n+2}, a_{n+3}, \ldots$ positive. In fact, $a_{n+1} = \lfloor \beta \rfloor \geq 1$ is positive by construction; here is where we used $\beta > 1$. We conclude that

$$\alpha = \frac{p_n \beta + p_{n-1}}{q_n \beta + q_{n-1}} = [a_0; a_1, \ldots, a_n, \beta] = [a_0; a_1, \ldots, a_n, a_{n+1}, a_{n+2}, \ldots].$$

By the uniqueness of the continued fraction (see Remark 1.51), we conclude that $(p_m, q_m) = (h_m, k_m)$ for $0 \leq m \leq n$, which completes the proof upon setting $m = n$. ■

1.3.6 Problems

Do at least ten points worth of the following exercises.

Problem 1.3.1 (2 points)**.** Work Exercise 1.44.

Problem 1.3.2 (2 points)**.** Compute the continued fraction of $\sqrt{23}$.

Problem 1.3.3 (3 points)**.** Work Exercise 1.60.

Problem 1.3.4 (3 points)**.** Let $a_0, a_1, a_2, \ldots$ be integers with $a_1, a_2, \ldots$ positive. Suppose that there exists an integer m such that $a_n = a_{n+m}$ for all n. Show that $[a_0; a_1, a_2, \ldots]$ is the root of a polynomial with integer coefficients and of degree two.

Problem 1.3.5 (4 points)**.** Find an irrational number $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ and integers h and k such that

$$\left| \alpha - \frac{h}{k} \right| < \frac{1}{k^2},$$

but h/k is not a continued fraction convergent of α.

Problem 1.3.6 (5 points)**.** Write (and submit) a Python program which takes as input an irrational number $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ and an index n and then outputs the nth coefficient a_n of the corresponding continued fraction $[a_0; a_1, a_2, \ldots]$ equal to α.

Problem 1.3.7 (8 points)**.** Let $\alpha \in \mathbb{R}$ be irrational and $[a_0; a_1, a_2, \ldots]$ its continued fraction expansion. Fix N sufficiently large. Suppose that among the first $1000N$ digits of the decimal expansion of α, the last $999N$ of them are all zeroes or all nines. Then there exists some $n \leq 5N$ so that $a_n > 10^{100N}$.

Problem 1.3.8 (2 points)**.** Use Problem 1.3.7 to conclude that for any sufficiently large N, the last $999N$ digits of the first $1000N$ decimal digits in the decimal expansion of $\sqrt{5}$ cannot be all zeroes or all nines.

1.4 Diophantine Approximation

Now that we have some experience with finding good rational approximations to real numbers, we are able to more firmly step foot into the field of Diophantine approximation. The content in this section is more intensive than in previous sections because it is essentially topics in Diophantine approximation.

1.4.1 Irrationality Measure

Fix an irrational number $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. From one perspective, the arc of the previous section was to go from knowing that there are infinitely rational numbers h/k such that

$$\left|\alpha - \frac{h}{k}\right| < \frac{1}{k}$$

to knowing that there are infinitely many rational numbers h/k such that

$$\left|\alpha - \frac{h}{k}\right| < \frac{1}{k^2}.$$

This is an amazing improvement: going from k to k^2 is a full exponent! But then we spent a lot of time improving the above result into

$$\left|\alpha - \frac{h}{k}\right| < \frac{1}{\sqrt{5}k^2},$$

which feels less significant because we are only improving by a constant. Of course, Example 1.58 established that we cannot do better than this in general, but for some real numbers, it will be possible. With this in mind, we take the following definition.

Definition 1.65 (irrationality measure)**.** Fix a real number $\alpha \in \mathbb{R}$. Then the *irrationality measure* $\mu(\alpha)$ of α is the least upper bound on the set of real numbers r such that

$$\left|\alpha - \frac{h}{k}\right| < \frac{1}{k^r}$$

for infinitely many rational numbers h/k with $k > 0$. Note that we allow $\mu(\alpha) = \infty$.

Remark 1.66. Note that there always is some real number r such that $\left|\alpha - \frac{h}{k}\right| < \frac{1}{k^r}$ for infinitely many rational numbers h/k, which makes the above definition make sense. Indeed, we may take $r = 1$. To see this, for any positive integer k, set $h \coloneqq \lfloor k\alpha \rfloor$ as in the previous section, so we find

$$\left|\alpha - \frac{h}{k}\right| = \frac{|k\alpha - \lfloor k\alpha \rfloor|}{k} < \frac{1}{k}.$$

So there are indeed infinitely many rational numbers h/k such that $\left|\alpha - \frac{h}{k}\right| < \frac{1}{k}$ as we let k vary.

Here are some early examples.

Example 1.67. Let α be a rational number. Then $\mu(\alpha) = 1$.

Solution. Remark 1.66 establishes $\mu(\alpha) \geq 1$. Further, for any $r > 1$, there are only finitely many rational numbers h/k such that

$$\left|\alpha - \frac{h}{k}\right| < \frac{1}{k^r}$$

by Proposition 1.45. Thus, $\mu(\alpha) \leq 1$, so the result follows. ■

Lemma 1.68. Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ be an irrational number. Then $\mu(\alpha) \geq 2$.

Proof. This follows directly from Corollary 1.53 upon unwinding. ■

Example 1.69. We have $\mu(\varphi) = 2$, where $\varphi = \frac{1+\sqrt{5}}{2}$.

Solution. Lemma 1.68 tells us that $\mu(\varphi) \geq 2$, so we just need to show that $\mu(\varphi) \leq 2$. It suffices to show that, for any $\varepsilon > 0$, there are only finitely many rational numbers h/k such that

$$\left|\varphi - \frac{h}{k}\right| < \frac{1}{k^{2+2\varepsilon}}.$$

Well, for sufficiently large k, we have $2k^{2+\varepsilon} < k^{2+2\varepsilon}$, so it is enough to show that there are finitely many rational numbers h/k such that

$$\left|\varphi - \frac{h}{k}\right| < \frac{1}{2k^{2+\varepsilon}}.$$

By Theorem 1.64, all such rational numbers h/k are continued fraction convergents. Thus, we take a moment to recall that $\{F_{n+2}/F_{n+1}\}_{n=0}^{\infty}$ are the continued fraction convergents of φ by Example 1.34, so it is enough to show that there are finitely many nonnegative integers n such that

$$\left|\varphi - \frac{F_{n+2}}{F_{n+1}}\right| < \frac{1}{2F_{n+1}^{2+\varepsilon}}.$$

However, Proposition 1.40 tells us that any nonnegative integer n has

$$\frac{1}{F_{n+1}(F_{n+1}+F_{n+2})} < \left|\varphi - \frac{F_{n+2}}{F_{n+1}}\right|,$$

so rearranging implies that it is enough to show there are only finitely many n with

$$2F_{n+1}^{\varepsilon} < \frac{F_{n+1}+F_{n+2}}{F_{n+1}} = 1 + \frac{F_{n+2}}{F_{n+1}}.$$

However, $F_{n+2}/F_{n+1} \to \varphi$ as $n \to \infty$, so the right-hand side is bounded while the left-hand side is not, so indeed there can be only finitely many n satisfying the above inequality. ■

Exercise 1.70. Show that $\mu\left(\sqrt{2}\right) = 2$.

Remark 1.71. It is not too hard to show that the continued fraction expansion for any quadratic irrational number α is eventually periodic, so the arguments of the previous two examples show that $\mu(\alpha) = 2$.

Example 1.72 (Liouville)**.** The real number

$$L := \sum_{k=0}^{\infty} \frac{1}{2^{k!}}$$

has $\mu(L) = +\infty$.

Proof. Quickly, note that the series converges because it is bounded above by $\sum_{k=0}^{\infty} 1/2^k = 2$. Now, for each natural n, define

$$L_n := \sum_{k=0}^{n} \frac{1}{2^{k!}}$$

to be the nth partial sum of L. Then L_n is a rational number with denominator $2^{n!}$, but

$$|L - L_n| = \sum_{k=n+1}^{\infty} \frac{1}{2^{k!}} < \sum_{k=(n+1)!}^{\infty} \frac{1}{2^k} = \frac{1}{2^{(n+1)!-1}}. \tag{1.2}$$

We are now ready to claim that $\mu(L) > r$ for any real number r. Indeed, for any real number r, we claim that there are infinitely many rational numbers h/k such that $|\alpha - h/k| < 1/k^r$. In fact, we claim that there are infinitely many n such that

$$|\alpha - L_n| < \frac{1}{2^{rn!}}.$$

Indeed, (1.2) implies that it is enough to show that

$$\frac{1}{2^{(n+1)!-1}} < \frac{1}{2^{rn!}}$$

for n sufficiently large, which is equivalent to $rn! < (n+1)! - 1$ for n sufficiently large, which is equivalent to $r < n + 1 - 1/n!$ for n sufficiently large, which is true. ■

Before continuing, we should note that essentially all real numbers α have irrationality measure 2. In particular, Example 1.69 is typical, and Example 1.72 is highly remarkable.

Proposition 1.73. Almost all real numbers $\alpha \in \mathbb{R}$ have $\mu(\alpha) = 2$. In other words, for any $\varepsilon > 0$, there is a countable collection of bounded intervals $\{(a_n, b_n)\}_{n=0}^{\infty}$ containing all $\alpha \in \mathbb{R}$ with $\mu(\alpha) = 2$ but

$$\sum_{n=0}^{\infty} (b_n - a_n) < \varepsilon.$$

Proof. Let S be the set of all $\alpha \in \mathbb{R}$ with $\mu(\alpha) \neq 2$. For brevity, let a subset $N \subseteq \mathbb{R}$ be a "null set" if and only if, for all $\varepsilon > 0$, there is a countable collection of bounded intervals $\{(a_n, b_n)\}_{n=0}^{\infty}$ containing N such that

$$\sum_{n=0}^{\infty} (b_n - a_n) < \varepsilon.$$

For example, we see that the union of countably many null sets is a null set by combining the countable collections $\{(a_n, b_n)\}_{n=0}^{\infty}$ together. Additionally, a point $\{x\}$ is a null set because x is covered by $(x - \varepsilon/2, x + \varepsilon/2)$ for any $\varepsilon > 0$. It follows from the previous two sentences that $\mathbb{Q}$ is a null set (it's a countable union of points), and thus it is enough to show that $S \setminus \mathbb{Q}$ is a null set.

Now, by Lemma 1.68, we see that $\alpha \in S \setminus \mathbb{Q}$ must have $\mu(\alpha) > 2$. For any real number $\varepsilon > 0$, we claim that

$$S_\varepsilon \coloneqq \{\alpha \in \mathbb{R} : \mu(\alpha) > 2 + \varepsilon\}$$

is a null set. By taking the union of $S = S_1 \cup S_{1/2} \cup S_{1/3} \cup \cdots$, it will follow that S is a null set. By taking another countable union, it is enough to show that $S_{\varepsilon,M} \coloneqq S_\varepsilon \cap [-M, M]$ is a null set for any $M > 0$. Well, any $\alpha \in S_\varepsilon$ has infinitely many rational numbers h/k such that

$$\left|\alpha - \frac{h}{k}\right| < \frac{1}{k^{2+\varepsilon}}.$$

In particular, $\alpha \in S_{\varepsilon,M}$ is contained in the set

$$S_{\varepsilon,M,K} \coloneqq \left\{\alpha \in [-M, M] : \left|\alpha - \frac{h}{k}\right| < \frac{1}{k^{2+\varepsilon}} \text{ for some } h \in \mathbb{Z} \text{ and } k \in \mathbb{Z} \cap [K, \infty)\right\}$$

for any $K > 1$. However, the above set is a countable union of small intervals: for each integer $k > K$, the set of relevant α have $2Mk+1$ options for h, and then each h has an interval of length $2/k^{2+\varepsilon}$ around it. The point is that $S_{\varepsilon,M,K}$ is covered by a countable union of intervals whose lengths total

$$\sum_{k>K} (2Mk+1) \cdot \frac{2}{k^{2+\varepsilon}} < \sum_{k>K} 3Mk \cdot \frac{2}{k^{2+\varepsilon}} = 6M \sum_{k>K} \frac{1}{k^{1+\varepsilon}}.$$

However, the series $\sum_{k=1}^{\infty} 1/k^{1+\varepsilon}$ converges, so the above length must go to zero as $K \to \infty$. Thus, for any $\delta > 0$, we can find K large enough so that $S_{\varepsilon,M,K}$ is covered by a countable union of intervals whose lengths sum to less than δ; this means that $S_{\varepsilon,M}$ is a null set, completing the proof. ■

1.4.2 Irrationality Measure via Continued Fractions

Example 1.69 has reminded us of the important fact that continued fraction convergents provide the best rational approximations. Thus, we might expect the irrationality measure to be controlled by the continued fraction convergents, which is indeed the case.

Lemma 1.74. Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ be an irrational number with continued fraction convergents $\{h_n/k_n\}_{n=0}^{\infty}$. Then

$$\mu(\alpha) = \limsup_{n\to\infty} \frac{-\log|\alpha - h_n/k_n|}{\log k_n}.$$

Proof. Let the $\limsup$ be L. Quickly, recall that Proposition 1.40 implies

$$\frac{-\log|\alpha - p_n/q_n|}{\log q_n} \geq \frac{-\log\left(1/q_n^2\right)}{\log q_n} = 2$$

for all n, so $L \geq 2$ has some lower bound.

For a given real number r, we claim that $\mu(\alpha) > r$ if and only if $L > r$, which complete the proof. Note that we may assume $r \geq 2$ because $\mu(\alpha) \geq 2$ by Lemma 1.68 and $L \geq 2$ already. Well, $\mu(\alpha) > r$ is equivalent to having some $\varepsilon > 0$ such that there are infinitely many rational numbers h/k such that

$$\left|\alpha - \frac{h}{k}\right| < \frac{1}{k^{r+\varepsilon}}.$$

Because $r \geq 2$, we see that $r + \varepsilon > 2$. Additionally, for k large enough, we have $2k^2 < 2k^{r+\varepsilon/2} < k^{r+\varepsilon}$, so all sufficiently large k must have h/k a continued fraction convergent. As such, this is equivalent to having infinitely many nonnegative integers n such that

$$\left|\alpha - \frac{h_n}{k_n}\right| < \frac{1}{k_n^{r+\varepsilon}},$$

or

$$\frac{-\log|\alpha - h_n/k_n|}{\log k_n} > r + \varepsilon.$$

This is now equivalent to $L \geq r + \varepsilon$ for our $\varepsilon > 0$, which is equivalent to $L > r$, completing the proof. ■

Proposition 1.75. Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ be an irrational number with continued fraction $[a_0; a_1, a_2, \ldots]$ and convergents $\{h_n/k_n\}_{n=0}^{\infty}$. Then

$$\mu(\alpha) = 1 + \limsup_{n\to\infty} \frac{\log k_{n+1}}{\log k_n} = 2 + \limsup_{n\to\infty} \frac{\log a_{n+1}}{\log k_n}.$$

Proof. We show the equalities separately.

- The left equality follows from Lemma 1.74. To see this, note that Proposition 1.40 implies

$$\frac{1}{2k_nk_n+1} < \frac{1}{k_n(k_n+k_{n+1})} < \left|\alpha - \frac{h_n}{k_n}\right| < \frac{1}{k_nk_{n+1}}$$

for any nonnegative integer n. Thus, Lemma 1.74 implies

$$\limsup_{n\to\infty} \frac{\log 2k_nk_{n+1}}{\log k_n} \geq \mu(\alpha) \geq \limsup_{n\to\infty} \frac{\log k_nk_{n+1}}{\log k_n},$$

which is equivalent to

$$1 + \limsup_{n\to\infty} \left(\frac{\log k_{n+1}}{\log k_n} + \frac{\log 2}{\log k_n}\right) \geq \mu(\alpha) \geq 1 + \limsup_{n\to\infty} \frac{\log k_{n+1}}{\log k_n},$$

For any $\varepsilon > 0$, there is N big enough so that $\log 2/\log k_n < \varepsilon$ for $n > N$, meaning

$$1 + \varepsilon + \limsup_{n\to\infty} \frac{\log k_{n+1}}{\log k_n} \geq \mu(\alpha) \geq 1 + \limsup_{n\to\infty} \frac{\log k_{n+1}}{\log k_n}.$$

Sending $\varepsilon \to 0^+$ completes the proof.

- The right equality follows from Proposition 1.32. To see this, recall from Proposition 1.32 that

$$k_{n+1} = a_{n+1}k_n + k_{n-1} = a_{n+1}k_n\left(1 + \frac{k_{n-1}}{a_{n+1}k_n}\right)$$

for $n \geq 1$, so

$$\limsup_{n\to\infty} \frac{\log k_{n+1}}{\log k_n} = 1 + \limsup_{n\to\infty} \left(\frac{\log a_{n+1}}{\log k_n} + \frac{\log\left(1 + \frac{k_{n-1}}{a_{n+1}k_n}\right)}{\log k_n}\right).$$

Notably, $0 \leq k_{n-1}/(a_{n+1}k_n) \leq 1$ for all $n \geq 1$, so we conclude that

$$\limsup_{n\to\infty} \frac{\log a_{n+1}}{\log k_n} \leq \mu(\alpha) - 2 \leq \limsup_{n\to\infty} \left(\frac{\log a_{n+1}}{\log k_n} + \frac{\log 2}{\log k_n}\right).$$

Now, for any $\varepsilon > 0$, we can find N so that $\log 2/\log k_n < \varepsilon$ for $n > N$, so sending $\varepsilon \to 0^+$ as above completes the proof. ■

Proposition 1.75 now gives us quite a bit of control over irrationality measure as long as we have control over the continued fraction. Here are some examples.

Example 1.76. Recall from Example 1.43 that $\varphi = [1; 1, 1, \ldots]$. Proposition 1.75 allows us to conclude that $\mu(\alpha) = 0$ immediately because $\log 1 = 0$.

Corollary 1.77. Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ be an irrational number with continued fraction $[a_0; a_1, a_2, \ldots]$. If there is a polynomial $f \in \mathbb{Z}[x]$ such that $a_n < f(n)$ for sufficiently large n, then $\mu(\alpha) = 2$.

Proof. By Proposition 1.75, it is enough to show that

$$\limsup_{n\to\infty} \frac{\log a_{n+1}}{\log k_n} = 0.$$

Now, because $a_n < f(n)$ for sufficiently large n, and $f(n) < n^{\deg f+1}$ for sufficiently large n, it is enough to show

$$\limsup_{n\to\infty} \frac{d\log n}{\log k_n} \leq 0$$

for any $d > 0$. Of course, we can now factor out the d and thus ignore it.

The main point is that $\{k_n\}_{n=0}^{\infty}$ increases at least exponentially. Explicitly, we claim is that $k_n \geq 1.5^{n-1}$ for any nonnegative n. This is by induction: certainly $k_0 = 1 \geq 1.5^{-1}$ and $k_1 = a_0 \geq 1.5^0$. Then for the induction, we see

$$k_{n+2} = a_{n+2}k_{n+1} + k_n \geq 1.5^n + 1.5^{n-1} > 1.5^{n-2},$$

where the last inequality holds because it rearranges to $1.5 + 1 > 1.5^2$, which is true.

Applying the main claim, we see that

$$0 \leq \limsup_{n\to\infty} \frac{\log n}{\log k_n} \leq \limsup_{n\to\infty} \frac{\log n}{1.5n} = 0,$$

which completes the proof. ∎

Corollary 1.78. Let r be any real number at least 2. Then there is an irrational $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ such that $\mu(\alpha) = r$.

Proof. For psychological reasons, we relegate $r = 2$ to Example 1.69. Otherwise, we may assume $r > 2$.

We construct α by infinite continued fraction $[a_0; a_1, a_2, \ldots]$, defining the a_n inductively using Proposition 1.32. Define $a_0 := 1$ and $a_1 := 2$ so that we have $k_0 := 1$ and $k_1 := 2$. Then for each $n \geq 1$, define $a_{n+1} := \lfloor k_n^{r-2} \rfloor$ and then $k_{n+1} := a_{n+1}k_n + k_{n-1}$ (as in Proposition 1.32). Then there is an integer sequence $\{h_n\}_{n=0}^{\infty}$ such that the rational numbers $\{h_n/k_n\}_{n=0}^{\infty}$ are the continued fraction convergents of $\alpha := [a_0; a_1, a_2, \ldots]$. By Corollary 1.48, we see that α is in fact irrational.

It remains to show that $\mu(\alpha) = r$. By Proposition 1.75, we would like to show that

$$r - 2 \stackrel{?}{=} \limsup_{n\to\infty} \frac{\log a_{n+1}}{\log k_n} = \limsup_{n\to\infty} \frac{\log \lfloor k_n^{r-2} \rfloor}{\log k_n}.$$

In fact, we claim that

$$\lim_{n\to\infty} \frac{\log \lfloor n^{r-2} \rfloor}{\log n} \stackrel{?}{=} r - 2,$$

which is good enough upon taking the limit along the subsequence $\{k_n\}_{n=0}^{\infty}$. Well, we see that

$$r - 2 = \frac{\log n^{r-2}}{\log n} \leq \frac{\log \lfloor n^{r-2} \rfloor}{\log n} \leq \frac{\log n^{r-2}}{\log n} + \frac{1}{\log n} = r - 2 + \frac{1}{\log n}$$

for any sufficiently large n, so we conclude upon taking $n \to \infty$. ∎

1.4.3 Algebraic Bounds on Irrationality Measure

One reason Diophantine approximation attracted the attention of number theorists is that one is able to use the condition that a number is algebraic in order to bound approximations. The prototypical and simplest result of this type is due to Liouville.

Definition 1.79 (algebraic, transcendental)**.** A nonzero complex number $\alpha \in \mathbb{C}$ is *algebraic of degree* d if and only if α is the root of an irreducible polynomial with rational coefficients and of degree d. We denote this degree d by $\deg \alpha$. If no such polynomial exists, then α is called *transcendental*.

Remark 1.80. Let's see that $\deg \alpha$ is well-defined: suppose that α is algebraic and hence the root of an irreducible polynomial f; by well-ordering, we may choose f to be of least degree. Then for any $g \in \mathbb{Q}[x]$ such that $g(\alpha) = 0$, we claim that f divides g (as polynomials in $\mathbb{Q}[x]$); taking g irreducible then forces $\deg f = \deg g$. Well, using the division algorithm for $\mathbb{Q}[x]$, we may write

$$g(x) = q(x)f(x) + r(x)$$

where $r = 0$ or $0 \leq \deg r < \deg f$. Plugging in α, we see that $r(\alpha) = 0$. Now, if $r \neq 0$, then we may factor r, and one of the irreducible factors will be irreducible, vanish at α, and have degree less than $\deg f$, contradicting the minimality of f. So instead $r = 0$, meaning f divides g.

Proposition 1.81 (Liouville)**.** Fix an algebraic real number $\alpha \in \mathbb{R}$ of degree $d \geq 2$. Then there exists a real number $\varepsilon > 0$ such that

$$\left|\alpha - \frac{h}{k}\right| > \frac{\varepsilon}{k^d}$$

for any rational number h/k with $k > 0$.

Proof. Let f be an irreducible polynomial with integer coefficients of degree $d \geq 2$ where $f(\alpha) = 0$. Namely, we may assume that f has integer coefficients by multiplying out a common denominator.

The main claim is that $|f(h/k)| \geq 1/k^d$. Indeed, $f(h/k) \neq 0$ because f may have no rational roots; explicitly, $f(h/k) = 0$ implies that $kx - h$ divides $f(x)$ by Remark 1.80, but f is irreducible, so this cannot be. But because f has integer coefficients, we may clear denominators to see $k^n f(h/k)$ is an integer. Explicitly, write $f(x) = \sum_{i=0}^{d} f_i x^i$ for integers $f_\bullet$, from which we find

$$k^n f\left(\frac{h}{k}\right) = \sum_{i=0}^{d} f_i h^i k^{d-i} \in \mathbb{Z}.$$

Thus, $|k^n f(h/k)| \geq 1$, and the claim follows.

To see how the above claim helps us, we note that

$$f(\alpha) - f\left(\frac{h}{k}\right) \approx \left(\alpha - \frac{h}{k}\right) f'(\alpha),$$

but the left-hand side has magnitude bounded below by $1/k^d$, so rearranging ought to give the result.

To make this rigorous, we begin by promising $\varepsilon < 1$ so that we might as well assume $|\alpha - h/k| < 1$. This allows us to make $\approx$ above into a genuine equality by using the Mean value theorem to find β between α and h/k such that

$$f(\alpha) - f\left(\frac{h}{k}\right) = \left(\alpha - \frac{h}{k}\right) f'(\beta).$$

Taking absolute values and using the main claim, we find

$$\left|\alpha - \frac{h}{k}\right| \geq \left|\frac{f(h/k)}{f'(\beta)}\right| \geq \frac{1}{|f'(\beta)|\, k^d}.$$

We now choose $\varepsilon > 0$ small enough so that $|f'(\beta_0)| < 1/\varepsilon$ for any $\beta_0 \in [\alpha - 1, \alpha + 1]$; such an upper bound exists because f' is a continuous function, and $[\alpha - 1, \alpha + 1]$ is compact. Because β is between α and h/k, and h/k is at most 1 away from α, this choice of ε completes the proof. ∎

Corollary 1.82. Fix an algebraic real number $\alpha \in \mathbb{R}$ of degree $d \geq 2$. Then $\mu(\alpha) \leq d$.

Proof. For any $\varepsilon > 0$, we show that there are only finitely many rational numbers h/k with $k > 0$ such that

$$\left|\alpha - \frac{h}{k}\right| < \frac{1}{k^{d+\varepsilon}},$$

which will show that $\mu(\alpha) < d+\varepsilon$ and hence complete the proof upon sending $\varepsilon \to 0^+$. Well, Proposition 1.81 grants some real number $\delta > 0$ such that

$$\left|\alpha - \frac{h}{k}\right| > \frac{\delta}{k^d}$$

for any rational number h/k with $k > 0$. Now, for sufficiently large $k > (1/\delta)^{1/\varepsilon}$, so $1/k^{d+\varepsilon} < \delta/k^d$, so indeed $|\alpha - h/k| < 1/k^{d+\varepsilon}$ is false for sufficiently large k. ■

Example 1.83. By Example 1.72, the number $L := \sum_{n=0}^{\infty} 2^{-n!}$ has $\mu(L) = +\infty$. Thus, Corollary 1.82 implies that L is transcendental.

Historically, examples of the type Example 1.83 were the first numbers proven to be transcendental.

Proposition 1.81 is not at all sharp. Using bivariate polynomials instead of single polynomials, Thue was able to sharpen Proposition 1.81 into the following.

Theorem 1.84 (Thue)**.** Fix an algebraic real number $\alpha \in \mathbb{R}$ of degree $d \geq 3$. Then for any $\varepsilon > 0$, there are only finitely many rational numbers h/k such that

$$\left|\alpha - \frac{h}{k}\right| < \frac{1}{k^{(d+1)/2+\varepsilon}}.$$

In other words, $\mu(\alpha) \leq (d+1)/2$.

The proof of Theorem 1.84 would add between five and ten pages to these notes, so we will not include it. However, it does not require anything much more serious than what we will cover in the remainder of this subsection.

Anyway, Theorem 1.84 is still not sharp. The following result is due to Roth and is work that earned Roth the Fields medal.

Theorem 1.85 (Roth)**.** Fix an algebraic real number $\alpha \in \mathbb{R}$. Then for any $\varepsilon > 0$, there are only finitely many rational numbers h/k such that

$$\left|\alpha - \frac{h}{k}\right| < \frac{1}{k^{2+\varepsilon}}.$$

In other words, $\mu(\alpha) = 2$.

It follows that all the numbers α we constructed in Corollary 1.78 with $\mu(\alpha) > 2$ were in fact transcendental! The proof of Theorem 1.85 would certainly take us too far afield, so we will not show it here.

Even though we will not prove Theorem 1.84, we will use it to the following result on Diophantine equations, as a means to reconnect with our roots.

Theorem 1.86 (Thue)**.** Let $f(x) = \sum_{k=0}^{d} f_k x^k \in \mathbb{Z}[x]$ be an irreducible polynomial of degree $d \geq 3$. For any $c \in \mathbb{Z}$, the equation

$$\sum_{k=0}^{d} f_k x^k y^{d-k} = c$$

has finitely many integer solutions (x, y).

Proof using Theorem 1.84. The idea is that x/y should be a good rational approximation to some real root of f, only finitely many of which should exist by Theorem 1.84.

Suppose for the sake of contradiction that there are infinitely many such solutions $\{(x_n, y_n)\}_{n=0}^{\infty}$. Our first task is to massage this sequence to converge (in some sense) to a root of f. For any given y, the equation we are solving is a polynomial in x equal to some constant, so there are only finitely many solutions. As such, we must have $|y_n| \to \infty$ as $n \to \infty$, so for example we may assume that $y_n \neq 0$ and that $\{|y_n|\}_{n=0}^{\infty}$ is a strictly increasing sequence. In this case, we see that

$$f\left(\frac{x}{y}\right) = \sum_{k=0}^{d} f_k \cdot \left(\frac{x}{y}\right)^k = y^{-d} \sum_{k=0}^{d} f_k x^k y^{d-k} = \frac{c}{y^d}$$

for each solution (x, y) with $y \neq 0$. Now, f is a polynomial of positive degree, so $|f(x)| \to \infty$ as $x \to \infty$, so having $|f(x/y)| = c/y^d \leq c$ forces x/y to live in some bounded interval $[-M, M]$. But then the infinite sequence $\{x_n/y_n\}_{n=0}^{\infty}$ must have a convergent subsequence, so we may assume that $\{x_n/y_n\}_{n=0}^{\infty}$ does in fact converge. Because $f(x_n/y_n) = c/y_n^d \to 0$ as $n \to \infty$, we see that $\{x_n/y_n\}_{n=0}^{\infty}$ converges to some real root α of f.

Our second task is to bound $|\alpha - x_n/y_n|$. Well, we may factor the irreducible polynomial f over $\mathbb{C}$ as

$$f(x) = f_d \prod_{k=1}^{d} (x - \alpha_k),$$

where $\{\alpha_1, \ldots, \alpha_d\}$ are the roots of f. We go ahead and rearrange the roots so that $\alpha_1 = \alpha$. As usual, note that these roots are disjoint for otherwise any double root would be a root shared by $f(x)$ and $f'(x)$, implying that $\gcd(f'(x), f(x))$ is a nontrivial factor of $f(x)$, thus violating irreducibility. We now see that

$$f_d \prod_{k=1}^{d} \left|\alpha_k - \frac{x_n}{y_n}\right| = f\left(\frac{x_n}{y_n}\right) = \frac{c}{|y_n|^d}$$

for each $n \geq 2$. We now isolate the $|\alpha - x_n/y_n|$ error term. For each $k \neq 2$, we find

$$\left|\alpha_k - \frac{x}{y}\right| > |\alpha_k - \alpha| - \left|\alpha - \frac{x}{y}\right|.$$

Now, by removing finitely many rational numbers from our sequence $\{x_n/y_n\}_{n=0}^{\infty}$, we may assume that $|\alpha - x_n/y_n|$ is less than $\frac{1}{2}|\alpha_k - \alpha|$ for each $k \neq 2$, which gives $|\alpha_k - x_n/y_n| > \frac{1}{2}|\alpha_k - \alpha|$. Thus,

$$\left|\alpha - \frac{x_n}{y_n}\right| < \underbrace{\frac{c}{f_d} \prod_{k=2}^{d} \frac{2}{|\alpha_k - \alpha|}}_{\delta :=} \cdot \frac{1}{|y_n|^d}.$$

Now, for $|y_n|$ sufficiently large, we will have $\delta / |y_n|^d < 1 / |y_n|^{(d+1)/2+1/4}$, so the infinitude of these rational approximations $\{x_n/y_n\}_{n=0}^{\infty}$ is now in direct contradiction with Theorem 1.84. ■

Example 1.87. The polynomial $f(x) := x^3 - 2$ is irreducible of degree 3. So Theorem 1.86 implies that $x^3 - 2y^3 = 10$ has only finitely many integer solutions (x, y). Indeed, there are at least two integer solutions $(2, -1)$ and $(4, 3)$.

1.4.4 e Is Transcendental

Thus far we have only constructed transcendental numbers by showing they have large irrationality measure, but we know from Proposition 1.73 that almost all real numbers have irrationality measure two. Because the set of algebraic numbers is countable (because the set of integer polynomials is countable), it follows that almost all transcendental numbers have irrationality measure two.

The goal of the next two subsections is to provide a single example of a transcendental number with irrationality measure two. In particular, we will show that e is transcendental and that $\mu(e) = 2$. In this subsection, we will show that e is transcendental; our exposition closely follows [Con]. To give us a flavor of the proof, we begin by showing that e is irrational.

Proposition 1.88. The real number e is irrational.

Proof. The main claim is that there is a sequence of rational numbers $\{p_n/q_n\}_{n=0}^{\infty}$ such that $|q_n e - p_n| \to 0$ and $q_n \to \infty$ as $n \to \infty$. Indeed, we set $q_n := n!$ and $p_n := \lfloor q_n e \rfloor$, which we compute by writing

$$n!e = n! \sum_{k=0}^{\infty} \frac{1}{k!} = \sum_{k=0}^{n} \frac{n!}{k!} + \sum_{n=k+1}^{\infty} \frac{n!}{k!},$$

so

$$n!e - \sum_{k=0}^{n} \frac{n!}{k!} = \sum_{k=n+1}^{\infty} \frac{1}{(n+1)(n+2)\cdots(k-1)k} < \sum_{k=n+1} \frac{1}{(n+1)^{k-n}} = \sum_{n=1}^{\infty} \frac{1}{(n+1)^k} = \frac{1/(n+1)}{1-1/(n+1)} \le 1,$$

meaning $p_n = \sum_{k=0}^{n} n!/k!$, and

$$|q_n e - p_n| < \left| \frac{1/(n+1)}{1-1/(n+1)} \right| = \frac{1}{n},$$

so indeed $|q_n e - p_n| \to 0$ as $n \to \infty$.

We now complete the proof. Suppose for the sake of contradiction that $e = p/q$ for some rational number p/q with $q > 0$ and $\gcd(p, q) = 1$. Then $|q_n p/q - p_n| \to 0$ as $n \to \infty$ by the above argument, so $|q_n p - p_n q| \to 0$. However, $q_n p - p_n q$ is an integer, so $q_n p = p_n q$ for n sufficiently large. But this does not make sense; for example, choosing n to be any prime, we see that $n \mid q_n$, so $n \mid p_n q$, but $p_n \equiv 1 \pmod{n}$ by definition of p_n, so $n \mid q$ instead. Thus, q must be larger than any prime, which is a contradiction. ∎

Remark 1.89. The above proof is in some sense the same argument as Proposition 1.45 applied to e; namely, we are using the close approximations p_n/q_n to e in order to derive a contradiction with the fact that all nonzero integers have magnitude at least 1. We have written it in the above manner to make the connection to the following transcendentality lemma clearer.

The crux of the above argument is the sequence of rational numbers $\{p_n/q_n\}_{n=0}^{\infty}$ such that $|q_n e - p_n| \to 0$ as $n \to \infty$. In order to show that e fails to be algebraic, the key is to find a way to simultaneously approximate not just e but also its powers. The following lemma explains how we will do this approximation.

Lemma 1.90. Fix a nonzero real number $\alpha \in \mathbb{R}$ and a positive integer d. Further, suppose that we have sequences of rational numbers $\{p_{1n}/q_n\}, \{p_{2n}/q_n\}, \ldots, \{p_{dn}/q_n\}$ satisfying the following.

(a) Approximation: for each k, we have $\left|q_n \alpha^k - p_{nk}\right| \to 0$ as $n \to \infty$.

(b) Technical: for each n, there is a common divisor g_n of the $p_{\bullet n}$ which is coprime to q_n but satisfies $g_n \to \infty$ as $n \to \infty$.

Then α is not the root of an irreducible polynomial in $\mathbb{Z}[x]$ of degree d.

Proof. Suppose for the sake of contradiction that $f(\alpha) = 0$ for some irreducible polynomial $f \in \mathbb{Z}[x]$ of degree d. To be explicit, write $f(x) = a_0 + a_1 x + \cdots + a_d x^d$ where $a_d \neq 0$. Note that $a_0 \neq 0$ because this would require $f(x) = x$, but $\alpha \neq 0$.

Now, the main idea is that p_{kn}/q_n should well-approximate α^k, so we go ahead and plug this into the "linear relation" $f(\alpha) = 0$. For any $n \geq 0$, we write

$$a_0 + \sum_{k=1}^{d} a_k \cdot \frac{p_{kn}}{q_n} = a_0 + \sum_{k=1}^{d} a_k \cdot \frac{p_{kn}}{q_n} - f(\alpha) = \sum_{k=1}^{d} a_k \left(\frac{p_{kn}}{q_n} - \alpha^k \right).$$

Clearing denominators, we find

$$q_n a_0 + \sum_{k=1}^{d} a_k p_{kn} = -\sum_{k=1}^{d} a_k \left(q_n \alpha^k - p_{kn} \right).$$

As $n \to \infty$, (a) tells us that the right-hand side goes to 0, so we must have

$$q_n a_0 = -\sum_{k=1}^{d} a_k p_{kn}$$

for n sufficiently large. However, this cannot be: g_n divides the right-hand side for all n, so $g_n \mid q_n a_0$, so $g_n \mid a_0$, which is a contradiction because a_0 is finite while $g_n \to \infty$ as $n \to \infty$. ■

It remains to construct these miraculous rational approximations p_{kn}/q_n of e^k. For this, we must use something about e; we will input the fact that $\frac{d}{dx} e^x = e^x$ into an integration by parts. To set up the relevant integration by parts, we will define

$$I_f(x) \coloneqq \sum_{k=0}^{\infty} f^{(k)}(x)$$

for any polynomial f. Notably, this sum is finite because the degree of f is finite. Here is our integration by parts result.

Lemma 1.91. For any polynomial f, we have

$$e^x \int_0^x e^{-t} f(t)\, dt = e^x I_f(0) - I_f(x).$$

Proof. Quickly, note that $I_f(x)$ is actually a finite sum because f is a polynomial. To get a taste of what is going on, we begin by writing the repeated integration by parts

$$\begin{aligned} \int_0^x e^{-t} f(t)\, dt &= f(0) - e^{-x} f(x) + \int_0^x e^{-t} f'(t)\, dt \\ &= (f(0) + f'(0)) - e^{-x} \left(f(x) + f'(x) \right) + \int_0^x e^{-t} f''(t)\, dt. \end{aligned}$$

This process continues. To make this rigorous, we define $I_f^m(x) \coloneqq \sum_{k=0}^{m} f^{(k)}(x)$, and we claim that

$$\int_0^x e^{-t} f(t)\, dt \stackrel{?}{=} I_f^m(0) - e^{-x} I_f^m(x) + \int_0^x e^{-t} f^{(m+1)}(t)\, dt$$

for any integer $m \geq -1$; the result will follow upon taking $m > \deg f$ so that $f^{(m)} = 0$. We show the claim by induction. At $m = -1$, there is nothing to say. For the inductive step, we note that integration by parts yields

$$I_f^m(0) - e^{-x} I_f^m(x) + \int_0^x e^{-t} f^{(m+1)}(t)\, dt = I_f^m(0) + f^{(m+1)}(0) - e^{-x} \left(I_f^m(x) + f^{(m+1)}(x) \right) + \int_0^x e^{-t} f^{(m+2)}(t)\, dt,$$

which is what we wanted upon rearranging and plugging into the inductive hypothesis. ■

We are now ready to prove the main result of this subsection.

Theorem 1.92. The real number e is transcendental.

Proof. Note that $e \neq 0$. We will use Lemma 1.90 show that e is not the root of any irreducible polynomial in $\mathbb{Z}[x]$ of degree d, for each $d \geq 1$. Thus, fixing some d, we need to construct the necessary sequences of rational numbers $\{p_{kn}/q_n\}$. For this, we use Lemma 1.91. We would like to approximate e^k, so we plug in $x = k$ to see that

$$e^k \int_0^k e^{-t} f(t)\, dt = e^k I_f(0) - I_f(k)$$

for any polynomial f. We would like the integral to be relatively small for each k between 0 and d, so we will set

$$f_n(t) \coloneqq t^{n-1}(t-1)^n(t-2)^n \cdots (t-d)^n$$

for $n \geq 1$. It is also important that f_r vanishes at $k \in \{1, 2, \ldots, d\}$ to a higher order than at 0. It is now tempting to directly set p_{kn}/q_n to be $I_{f_n}(k)/I_{f_n}(0)$, but we will want to use the high vanishing of f_n in order to factor out from p_{kn} and q_n beforehand.

Indeed, for each $k \in \{0, 1, \ldots, d\}$, we have the Taylor expansion

$$f_n(t+k) = \sum_{\ell=0}^{\infty} \frac{f^{(\ell)}(k)}{\ell!} \cdot t^\ell,$$

but these coefficients must all be integers, so we conclude that $\ell! \mid f_n^{(\ell)}(k)$ for all nonnegative integers ℓ. At $k = 0$, we actually have $f_n^{(\ell)}(0) = 0$ for $0 \leq \ell \leq n-1$; and for $k \in \{1, 2, \ldots, d\}$, we have $f_n^{(\ell)}(k) = 0$ for $0 \leq \ell \leq n$. Thus, $I_{f_n}(k)$ is divisible by $(n-1)!$ for each k, but it is divisible by $n!$ for each $k > 0$ while

$$I_{f_n}(0) \equiv f^{(n-1)}(0) \equiv (-1)^{nd} d! \pmod{n!} \tag{1.3}$$

because the higher-order terms are $0 \pmod{n!}$.

With this in mind, we set $p_{kn} \coloneqq I_{f_n}(k)/(n-1)!$ and $q_n \coloneqq I_{f_n}(0)/(n-1)!$ for each nonnegative integer n. We now check (a) and (b) of Lemma 1.90, which will complete the proof.

(a) We compute

$$\begin{aligned} \left|q_n e^k - p_{nk}\right| &\leq \frac{e^k}{(n-1)!} \int_0^k e^{-t} \left|f_n(t)\right| \, dt \\ &= \frac{e^k}{(n-1)!} \int_0^k e^{-t} \left|t\right|^{n-1} \left|t-1\right|^n \left|t-2\right|^n \cdots \left|t-d\right|^n \, dt \\ &\leq \frac{e^k}{(n-1)!} \cdot d^{n-1+dn} \int_0^k e^{-t}\, dt. \end{aligned}$$

Now, $\int_0^k e^{-t}\, dt < \int_0^\infty e^{-t}\, dt = 1$, so we have the bound

$$\left|q_n e^k - p_{nk}\right| < \frac{e^k}{d} \cdot \frac{\left(d^{d+1}\right)^n}{(n-1)!}.$$

The right-hand side goes to 0 as $n \to \infty$, so the left-hand side must also.

(b) For this check, we actually want to use a subsequence of the rationals we chose. The common divisor will be $g_n \coloneqq n$, which we know divides each $p_{kn} = I_{f_n}(k)/(n-1)!$ because $I_{f_n}(k)$ is divisible by $n!$. However, we must verify that there are infinitely many n such that n is relatively prime to q_n. Well, (1.3) implies that it is enough for n to be relatively prime to $d!$, so we may take $n(m) \coloneqq 1 + md!$ and then take our rationals to be $\left\{p_{k,n(m)}/q_{n(m)}\right\}$. This completes the proof. ∎

1.4.5 The Continued Fraction of e

In this subsection, we compute the continued fraction expansion of e and then use it to show that $\mu(e) = 2$. Our exposition in this subsection follows [Old70]. We are going to prove that

$$e \stackrel{?}{=} [2; 1, 2, 1, 1, 4, 1, 1, 6, 1, \ldots, 1, 2m, 1, \ldots].$$

This continued fraction naturally comes in threes, so it will actually be easier to show the related continued fraction

$$\frac{e+1}{e-1} = [2; 6, 10, 14, \ldots, 4m+2, \ldots].$$

Nonetheless, the main part of our story will unsurprisingly be focused on trying to come up with good rational approximations for e. Anyway, let's jump into a proof.

Proposition 1.93. We have

$$\frac{e+1}{e-1} = [2; 6, 10, 14, \ldots, 4m+2, \ldots].$$

Proof. For clarity, we proceed in steps.

1. We produce reasonably good rational approximations p_n/q_n (for nonnegative integers n) to e. By Lemma 1.91, we have
$$e \int_0^1 e^{-t} f(t)\, dt = eI_f(0) - I_f(1)$$
for any polynomial f. We would like to make the integral small in order to produce a good rational approximation of e, so we will take our polynomial to be $f_n(t) \coloneqq t^n(t-1)^n$. Arguing as in Theorem 1.92, we see that $I_{f_n}(0)$ and $I_{f_n}(1)$ are integers divisible by $n!$. Indeed, the Taylor expansion
$$f_n(t+k) = \sum_{\ell=0}^{\infty} \frac{f^{(\ell)}(k)}{\ell!} \cdot t^\ell$$
establishes that $f_n^{(\ell)}(k)$ is an integer divisible by $\ell!$ for any $\ell \geq 0$. However, $f_n^{(\ell)}(k) = 0$ for $k \in \{0, 1\}$ and $0 \leq \ell \leq n$ by construction of f_n, so we conclude that $I_{f_n}(k)$ is divisible by $n!$ for $k \in \{0, 1\}$ because all nonzero terms of the sum
$$I_{f_n}(k) = \sum_{\ell=0}^{\infty} f_n^{(\ell)}(k)$$
are divisible by $n!$.

   Thus, we define $q_n \coloneqq I_{f_n}(0)/n!$ and $p_n \coloneqq I_{f_n}(1)/n!$. To verify that p_n/q_n is in fact a good rational approximation to e, we write
$$\begin{aligned} |q_n e - p_n| &\leq \frac{e}{n!} \int_0^1 e^{-t} \left|f_n(t)\right| \, dt \\ &= \frac{e}{n!} \int_0^1 e^{-t} \left|t(t-1)\right|^n \, dt \\ &< \frac{e}{n!} \int_0^\infty e^{-t} \, dt \\ &= \frac{e}{n!}. \end{aligned}$$

2. We produce a recurrence relation for the $\{p_n\}$ and $\{q_n\}$. This will arise purely formally by manipulating the integrals
$$J_{a,b} \coloneqq \int_0^1 e^{-t} t^a (t-1)^b \, dt.$$

We have two "moves": on one hand, integration by parts shows

$$J_{a,b} = \int_0^1 e^{-t}t^a(t-1)^b\,dt = a\int_0^1 e^{-t}t^{a-1}(t-1)^b\,dt + b\int_0^1 e^{-t}t^a(t-1)^{b-1}\,dt = aJ_{a-1,b} + bJ_{a,b-1} \tag{1.4}$$

for $a, b \geq 1$, and the identity $(t-1)^b = t(t-1)^{b-1} - (t-1)^{b-1}$ shows

$$J_{a,b} = \int_0^1 e^{-t}t^a(t-1)^b\,dt = \int_0^1 e^{-t}t^{a+1}(t-1)^{b-1}\,dt - \int_0^1 e^{-t}t^a(t-1)^{b-1}\,dt = J_{a+1,b-1} - J_{a,b-1} \tag{1.5}$$

for $a \geq 0$ and $b \geq 1$. Now, the main claim of this step is that

$$J_{n,n} \stackrel{?}{=} 2n(2n-1)J_{n-1,n-1} + n(n-1)J_{n-2,n-2} \tag{1.6}$$

for $n \geq 2$. We will prove this using (1.4) and (1.5) repeatedly. Getting started, we write

$$\begin{aligned} J_{n,n} &= nJ_{n-1,n} + nJ_{n,n-1} && (1.4)\\ &= nJ_{n-1,n} + n(J_{n-1,n} + J_{n-1,n-1}) && (1.5)\\ &= 2nJ_{n-1,n} + nJ_{n-1,n-1}. \end{aligned}$$

The relation

$$J_{n,n} = 2nJ_{n-1,n} + nJ_{n-1,n-1} \tag{1.7}$$

will be helpful again in a moment. Anyway, we now continue, writing

$$\begin{aligned} J_{n,n} &= nJ_{n-1,n-1} + 2n((n-1)J_{n-2,n} + nJ_{n-1,n-1}) && (1.4)\\ &= \left(2n^2+n\right)J_{n-1,n-1} + \left(2n^2-2n\right)J_{n-2,n}\\ &= \left(2n^2+n\right)J_{n-1,n-1} + \left(2n^2-2n\right)(J_{n-1,n-1} - J_{n-2,n-1}) && (1.5)\\ &= \left(4n^2-n\right)J_{n-1,n-1} - 2n\,(n-1)\,J_{n-2,n-1}\\ &= \left(4n^2-n\right)J_{n-1,n-1} - n(J_{n-1,n-1} - (n-1)J_{n-2,n-2}) && (1.7)\\ &= 2n(2n-1)J_{n-1,n-1} + n(n-1)J_{n-2,n-2}, \end{aligned}$$

which is precisely (1.6).

We now conclude this step. Note that

$$q_ne - p_n = \frac{e}{n!}\int_0^1 e^{-t}t^n(t-1)^n\,dt = \frac{e}{n!}\cdot J_{n,n},$$

so the recurrence (1.6) implies that

$$q_ne - p_n = 2(2n-1)(q_{n-1}e - p_{n-1}) + (q_{n-2}e - p_{n-2})$$

for $n \geq 2$. Because e is irrational (by Proposition 1.88), collecting terms to put all es on one side and all integers on the other, we produce the system of recurrences

$$\begin{cases} p_{n+1} = 2(2n+1)p_n + p_{n-1}, \\ q_{n+1} = 2(2n+1)q_n + q_{n-1}, \end{cases} \tag{1.8}$$

for $n \geq 1$. These recurrences essentially explain why the desired continued fraction expansion features $4m+2$.

3. We take a linear combination of the $\{p_n\}$ and $\{q_n\}$ to produce continued fraction convergents. To see why we must do this, we begin by computing (p_0, q_0) and (p_1, q_1). On one hand, $f_0(t) = 1$, so $I_{f_0}(0) = I_{f_0}(1) = 1$, so $(p_0, q_0) = (1, 1)$. On the other hand, $f_1(t) = t(t-1) = t^2 - t$ had $f_1'(t) = 2t-1$ and $f''(t) = 2$, so $I_{f_1}(0) = 1$ and $I_{f_1}(1) = 3$, so $(p_1, q_1) = (3, 1)$.

The number $p_0/q_0 = 1$ is not a continued fraction convergent of e, so there is no way of shifting our sequences directly in order to produce the continued fraction for e. However, if one sets $(h_{-2}, k_{-2}) = (0, 1)$ and $(h_n, k_n) \coloneqq ((p_{n+1} + q_{n+1})/2, (p_{n+1} - q_{n+1})/2)$ for each $n \geq -1$, then we see

$$\begin{cases} h_n = 2(2n+1)h_{n-1} + h_{n-2}, \\ k_n = 2(2n+1)k_{n-1} + k_{n-2}, \end{cases}$$

for $n \geq -2$ by an explicit computation at $n = -2$ and (1.8) for $n \geq -1$. Thus, by Proposition 1.32, we see

$$\frac{h_n}{k_n} = [2; 6, 10, 14, \ldots, 4n+2]$$

for each $n \geq 0$.

4. We complete the proof. It remains to show that $h_n/k_n \to (e+1)/(e-1)$ as $n \to \infty$. The bound

$$|q_n e - p_n| < \frac{e}{n!}$$

now reads

$$|(h_n - k_n)e - (h_n + k_n)| < \frac{e}{(n+1)!},$$

which rearranges to

$$\left| \frac{e+1}{e-1} - \frac{h_n}{k_n} \right| < \frac{e}{(e-1)(n+1)!k_n}.$$

Sending $n \to \infty$ completes the proof. ■

To produce the continued fraction for e, we need to be able to manipulate continued fractions. We will want the following.

Lemma 1.94. Fix any positive real numbers $a, b, c \in \mathbb{R}$ with $a, b \geq 1$. Then $2/[a; b, c] = 1/[(a-1)/2, 1, 1 + 2/[b-1, a]]$.

Proof. The lower bounds on $a, b, c \in \mathbb{R}$ are merely there to ensure we have no division by zero problems. For example, we have ensured $[b-1; a] > 0$ currently. Anyway, unwrapping, we are trying to show

$$\frac{2}{a + \dfrac{1}{b + \dfrac{1}{c}}} \stackrel{?}{=} \frac{1}{\dfrac{a-1}{2} + \dfrac{1}{1 + \dfrac{1}{1 + \dfrac{2}{b - 1 + \dfrac{1}{c}}}}}.$$

This is purely formal. Taking reciprocals, we are trying to show

$$a + \frac{1}{b + \dfrac{1}{c}} \stackrel{?}{=} (a-1) + \frac{2}{1 + \dfrac{1}{1 + \dfrac{2}{b - 1 + \dfrac{1}{c}}}}.$$

Now, the fraction on the right-hand side is

$$\frac{2}{1 + \dfrac{1}{1 + \dfrac{2c}{bc - c + 1}}} = \frac{2}{1 + \dfrac{bc - c + 1}{bc + c + 1}} = \frac{bc + c + 1}{bc + 1} = 1 + \frac{c}{bc + 1} = 1 + \frac{1}{b + \dfrac{1}{c}},$$

which completes the proof upon plugging in to the previous equation. ■

Theorem 1.95. We have
$$e = [2; 1, 2, 1, 1, 4, 1, 1, 6, 1, \ldots, 1, 2m, 1, \ldots].$$

Proof. Subtracting one and taking the reciprocal from Proposition 1.93, we find
$$\frac{e-1}{2} = [0; 1, 6, 10, 14, \ldots, 4m+2, \ldots].$$
Rearranging, we find
$$e = 1 + 2/[1; 6, 10, 14, \ldots, 4m+2, \ldots].$$
Beginning our translation, we use Lemma 1.94 to see that this is
$$e = 1 + 1/[0; 1, 1 + 2/[5; 10, 14, 18, \ldots]] = [1; 0, 1, 1 + 2/[5; 10, 14, 18, \ldots]].$$
More generally, we claim that
$$e \stackrel{?}{=} [1; 0, 1, 1, 2, 1, \ldots, 1, 2m, 1, 1 + 2/[4m+5; 4m+10, 4m+14, 4m+18, \ldots]]$$
for any $m \geq 0$. We just showed the $m = 0$ case. For the induction, we use Lemma 1.94 to find
$$\begin{aligned} e &= [1; 0, 1, 1, 2, 1, \ldots, 1, 2m, 1, 1 + 2/[4m+5; 4m+10, 4m+14, 4m+18, \ldots]] \\ &= [1; 0, 1, 1, 2, 1, \ldots, 1, 2m, 1, 1 + 1/[2m+1; 1, 1 + 2/[4m+9, 4m+14, 4m+18, \ldots]]] \\ &= [1; 0, 1, 1, 2, 1, \ldots, 1, 2m, 1, 1, 2m+2, 1, 1 + 2/[4m+9, 4m+14, 4m+18, \ldots]]. \end{aligned}$$
Sending $m \to \infty$ and adjusting the start of the continued fraction completes the proof. Formally, one should justify why sending $m \to \infty$ makes the continued fraction converge, but this holds essentially by the argument of Proposition 1.40 because the last coefficient of the continued fraction above is always bigger than one and hence unable to cause problems with convergence. ■

Corollary 1.96. We have $\mu(e) = 2$.

Proof. This follows directly from plugging in Theorem 1.95 into Corollary 1.77. For example, the polynomial $f(n) = n + 3$ will do the trick. ■

1.4.6 Problems

Do ten points worth of the following exercises.

Problem 1.4.1 (1 points)**.** Compute the first five continued fraction convergents of e.

Problem 1.4.2 (2 points)**.** Without appeal to results unproven in these notes, work Exercise 1.70.

Problem 1.4.3 (3 points)**.** Show that the real number
$$\sum_{n=0}^{\infty} \frac{n}{10^{n!}}$$
is transcendental.

Problem 1.4.4 (3 points)**.** Compute

$$\int_0^1 e^{-t} t^5 (t-1)^4 \, dt.$$

Problem 1.4.5 (3 points)**.** For any real number $r \geq 2$, show that the set

$$S_r \coloneqq \{\alpha \in \mathbb{R} : \mu(\alpha) = r\}$$

is dense in $\mathbb{R}$. In other words, for any real number $x \in \mathbb{R}$ and positive $\varepsilon > 0$, show that there is some $\alpha \in S_r$ such that $|x - \alpha| < \varepsilon$.

Problem 1.4.6 (4 points)**.** Consider the real number

$$L = \sum_{k=0}^{\infty} \frac{1}{2^{3^k}}.$$

Show that $\mu(L) \geq 3$.

Problem 1.4.7 (5 points)**.** For $|y| > 100$, show that any integer pair (x, y) such that $x^3 - 2y^3 = 10$ must have x/y be a continued fraction convergent of $\sqrt[3]{2}$. Using Sage, show that there are no solutions aside $(x, y) \in \{(2, -1), (4, 3)\}$ with $|x|, |y| < 10^{100}$. Please submit the program.

Problem 1.4.8 (10 points)**.** Adapt the proof of Proposition 1.93 to show that

$$\frac{e^{2/k} + 1}{e^{2/k} - 1} = [k; 3k, 5k, \ldots]$$

for any integer $k \geq 2$. You may find [Old70] helpful.

THEME 2
QUADRATIC EQUATIONS: UNITS

attempts to change the terminology introduced by Euler have always proved futile.

—H. W. Lenstra Jr., [Jr02]

2.1 Pell Equations

The goal of the present section is to discuss equations of the form

$$ax^2 + bxy + cy^2 = d$$

where $a, b, c, d \in \mathbb{Z}$. Completing the square and completing the denominator, we might as well solve

$$ax^2 + by^2 = c$$

where $a, b, c \in \mathbb{Z}$. Multiplying through by a, we are trying to solve

$$(ax)^2 + (ab)y^2 = ac,$$

so we may as well try to find integer solutions to the equation

$$x^2 - dy^2 = c$$

where $d, c \in \mathbb{Z}$. If $d < 0$, then $x^2 - dy^2 = c$ must have $|x| \leq \sqrt{c}$ and $|y| < \sqrt{c/d}$, so solving this equation can be done via a finite computation. Otherwise, $d > 0$. From here, it turns out that we can produce much of the internal structure of the solutions by limiting our view to $|c| = 1$, which we will call "Pell equations" for now (though we will want to expand our definition). This will remain our focus for the majority of this section.

2.1.1 Pell Equations via Elementary Methods

Shortly we are going to begin discussing real quadratic fields and their connections to Pell equations, but it is worthwhile to be aware that one can make purely elementary arguments to solve these equations. Let's see a few examples and feel the wonder.

Remark 2.1. The following examples, suitably transformed, can also be seen as a form of "Vieta jumping." We will not bother to explain what Vieta jumping is, but those who do may find Problem 2.1.3 compelling.

Example 2.2. Define the sequence of ordered pairs of nonnegative integers $\{(x_n, y_n)\}_{n=0}^{\infty}$ recursively by $(x_0, y_0) := (1, 0)$ and

$$(x_{n+1}, y_{n+1}) := (2x_n + 3y_n, x_n + 2y_n)$$

for any $n \geq 0$. Then for any pair of nonnegative integers (x, y) such that $x^2 - 3y^2 = 1$, we have $(x, y) = (x_n, y_n)$ for some nonnegative integer n, and (x_n, y_n) is a solution for each n.

Solution. We have two claims to show, so we will show them separately. The main characters of this solution are the linear transformation $f_{\pm} \colon \mathbb{Z}^2 \to \mathbb{Z}^2$ given by $f_{\pm}(x, y) := (2x \pm 3y, \pm x + 2y)$ where $\pm$ is some sign; in particular, $f_+(x_n, y_n) = (x_{n+1}, y_{n+1})$.

1. We show that $x_n^2 - 3y_n^2 = 1$ for all nonnegative integers n. We induct on n. At $n = 0$, we are saying $1^2 - 3 \cdot 0^2 = 1$, which is true. For the inductive step, we show that $f_{\pm}(x, y)$ is a solution of (x, y) is for any sign $+$ or $-$. Well, if $x^2 - 3y^2 = 1$, then we compute

$$(2x \pm 3y)^2 - 3(\pm x + 2y)^2 = \left(4x^2 \pm 12xy + 9y^2\right) - 3\left(x^2 \pm 4xy + 4y^2\right) = x^2 - 3y^2 = 1,$$

   so $f(x, y)$ is also a solution.

2. As an intermediate step, we check that f_+ and f_- are inverse functions. Well, we compute

$$f_{\pm}(f_{\mp}(x, y)) = f(2x \mp 3y, \mp x + 2y) = (2(2x \mp 3y) \pm 3(\mp x + 2y), \pm(2x \mp 3y) + 2(\mp x + 2y)) = (x, y)$$

   for any arrangement of signs.

3. Lastly, fix a solution (x, y) of $x^2 - 3y^2 = 1$. We would like to show that $(x, y) = (x_n, y_n)$ for some n, which is equivalent to $(x, y) = f_+^n(1, 0)$, or $f_-^n(x, y) = (1, 0)$ for some n. Well, let n be the largest nonnegative integer such that both entries of $(x', y') := f_-^n(x, y)$ are both nonnegative integers; we claim that $(x', y') = (1, 0)$, which will complete the proof. The first step establishes that $(x')^2 - 3(y')^2 = 1$, and we know that one of the coordinates of $f_-(x', y')$ must fail to be a nonnegative integer. Thus, we either have $2x' - 3y' < 0$ or $-x' + 2y' < 0$, so $2x' < 3y'$ or $2y' < x'$.

   If $2x' < 3y'$, then

$$1 = (x')^2 - 3(y')^2 < \left(\frac{3}{2} \cdot y'\right)^2 - 3(y')^2 < 0,$$

   so this cannot be. Thus, we must instead have $x' > 2y'$, from which we find

$$1 = (x')^2 - 3(y')^2 > (2y')^3 - 3(y')^2 = (y')^2,$$

   so $y' = 0$, so $(x', y') = (1, 0)$. ∎

Example 2.3. Define the sequence of ordered pairs of nonnegative integers $\{(x_n, y_n)\}_{n=0}^{\infty}$ recursively by $(x_n, y_n) := (1, 0)$ and

$$(x_{n+1}, y_{n+1}) := (x_n + 2y_n, x_n + y_n)$$

for any $n \geq 0$. Then for any pair of nonnegative integers (x, y) such that $x^2 - 2y^2 = \pm 1$, we have $(x, y) = (x_n, y_n)$ for some nonnegative integer n, and (x_n, y_n) is a solution for each n. In fact, $x_n^2 - 2y_n^2 = (-1)^n$ for each n.

Solution. Again, we have two claims to show, and we will show them separately. As before, the main characters of this solution are the linear transformations $f_\pm\colon \mathbb{Z}^2 \to \mathbb{Z}^2$ given by $f_\pm(x,y) \coloneqq (\pm x+2y, x\pm y)$where $\pm$ is some sign; in particular, $f_+(x_n,y_n) = (x_{n+1},y_{n+1})$.

1. We show that $x_n^2 - 2y_n^2 = (\pm 1)^n$ for each nonnegative integer n. We proceed by induction. At $n=0$, we are saying that $1^2 - 2\cdot 0^2 = 1$. For the inductive step, we suppose $x^2-2y^2 = (-1)^n$ and show that $(x',y') \coloneqq f_\pm(x,y)$ has $(x')^2 - 2(y')^2 = -(-1)^{n+1}$. Indeed, we compute
$$(\pm x+2y)^2 - 2(x\pm y)^2 = \left(x^2 \pm 4xy + 4y^2\right) - 2\left(x^2 \pm 2xy + y^2\right) = -\left(x^2-2y^2\right) = (-1)^{n+1}.$$

2. We check that f_+ and f_- are inverse functions. Well, we compute
$$f_\pm(f_\mp(x,y)) = f_\pm(\mp x+2y, x\mp y) = (\pm(\mp x+2y)+2(x\mp y), (\mp x+2y)\pm(x\mp y)) = (x,y)$$
for any arrangement of signs.

3. Lastly, fix a solution (x,y) of $x^2-2y^2 = \pm 1$. We would like to show that $(x,y) = (x_n,y_n)$ for some $n\geq 0$, which is equivalent to $(x,y) = f_+^n(1,0)$, or $f_-^n(x,y) = (1,0)$ for some n. Well, let n be the largest nonnegative integer such that both entries of $(x',y') \coloneqq f_-^n(x,y)$ are both nonnegative integers; we claim that $(x',y') = (1,0)$, which will complete the proof.

   The first step gives $(x')^2 - 2(y')^2 = \pm 1$ still, and because a coordinate of $f_-(x',y')$ must be a negative integer, we have either $2y' < x'$ or $x' < y'$. On one hand, if $x' < y'$, then
$$\pm 1 = (x')^2 - 2(y')^2 < (y')^2 - 2(y')^2 = -(y')^2,$$
   so we must have $(y')^2 < \mp 1 \leq 1$, so $y' = 0$, which forces $x' < 0$, which makes no sense. On the other hand, if $2y' < x'$, then
$$\pm 1 = (x')^2 - 2(y')^2 > (2y')^2 - 2(y')^2 = 2(y')^2,$$
   so $y' = 0$ is still forced, from which we must have $x' = 1$, so $(x',y') = (1,0)$. ∎

Exercise 2.4. Define the sequence of ordered pairs of nonnegative integers $\{(x_n,y_n)\}_{n=0}^\infty$ recursively by $(x_n,y_n) \coloneqq (1,0)$ and
$$(x_{n+1},y_{n+1}) \coloneqq (3x_n+4y_n, 2x_n+3y_n)$$
for any $n\geq 0$. Then for any pair of nonnegative integers (x,y) such that $x^2-2y^2 = 1$, show that $(x,y) = (x_n,y_n)$ for some nonnegative integer n. Describe the solutions to $x^2-2y^2 = -1$ similarly.

One might look at Example 2.3 and wonder what all the fuss with $(-1)^n$ is, for it is a perfectly reasonable question to look for solutions to $x^2-2y^2=1$ on its own, as shown by the above exercise. However, the recursion $(x,y)\mapsto(3x+4y,2x+3y)$ is in some sense "more complicated than it has to be" both in the sense that the coefficients are larger than the recursion in Example 2.3 and also in the sense that this recursion is simply the recursion in Example 2.3 applied twice:
$$(x,y)\mapsto(x+2y,x+y)\mapsto(3x+4y,2x+3y).$$
As such, it turns out that the "correct" thing to do is in fact to look at solutions to $x^2-2y^2=\pm 1$ and then correct at the end to look at solutions to $x^2-2y^2=1$. The reason why is explained somewhat but not completely in section 2.1.2. A complete explanation will have to wait for the next section.

The following example provides the extreme end of trying to make our recursions as simple as possible.

Example 2.5. Define the sequence of ordered pairs of nonnegative integers $\{(x_n,y_n)\}_{n=0}^\infty$ recursively by $(x_0,y_0) \coloneqq (2,0)$ and
$$(x_{n+1},y_{n+1}) \coloneqq \left(\frac{x_n+5y_n}{2}, \frac{x_{n+1}+y_{n+1}}{2}\right)$$
for any $n\geq 0$. Then for any pair of nonnegative integers (x,y) such that $x^2-5y^2 = \pm 4$, we have $(x,y) = (x_n,y_n)$ for some nonnegative integer n, and (x_n,y_n) is a solution for each n. In fact, $x_n^2 + x_ny_n - y_n^2 = (-1)^n\cdot 4$ for each n.

Solution. The proof is similar to the previous two ones. We define the linear transformations $f_\pm\colon \mathbb{Z}^2 \to \mathbb{Z}^2$ by $f_\pm(x,y) \coloneqq \frac{1}{2}(\pm x + 5y, x \pm y)$ so that $f_+(x_n, y_n) = (x_{n+1}, y_{n+1})$.

1. We show that $f_\pm(x,y)$ is a pair of integers of the same parity whenever (x,y) is a pair of integers of the same parity. This verifies that $\{(x_n,y_n)\}_{n=0}^\infty$ is in fact a sequence of integers (by induction) because $(x_0,y_0) = (2,0)$ is a pair of integers of the same parity. Well, if x and y are both the same parity, then $\pm x + 5y$ and $x \pm y$ are both even, so $f_\pm(x,y)$ is a pair of integers. To see that $\frac{1}{2}(\pm x + 5y)$ and $\frac{1}{2}(x \pm y)$ are both the same parity, we note that
$$\frac{\pm x + 5y}{2} \mp \frac{x \pm y}{2} = 2y \equiv 0 \pmod 2.$$

2. We show that $x_n^2 - 5y_n^2 = (-1)^n \cdot 4$ for each $n \geq 0$. For $n = 0$, we are saying that $2^2 - 5 \cdot 0^2 = 4$, which is true. For the inductive step, we more generally show that $(x')^2 - 5(y')^2 = \mp 4$ when $(x',y') = f_\pm(x,y)$ for $x^2 - 5y^2 = \pm 4$. Well, suppose $x^2 - 5y^2 = \pm 4$, and we compute
$$\left(\frac{\pm x + 5y}{2}\right)^2 - 5\left(\frac{x \pm y}{2}\right)^2 = \frac{x^2 \pm 10xy + 25y^2}{4} - 5 \cdot \frac{x^2 \pm 2xy + y^2}{4} = -\left(x^2 - 5y^2\right) = \mp 4.$$

3. We check that f_+ and f_- are inverse functions. Well, we compute
$$\begin{aligned} f_\pm(f_\mp(x,y)) &= f_\pm\left(\frac{\mp x + 5y}{2}, \frac{x \mp y}{2}\right) \\ &= \frac{1}{4}(\pm(\mp x + 5y) + 5(x \mp y), (\mp x + 5y) \pm (\mp x + y)) \\ &= (x,y). \end{aligned}$$

4. Lastly, fix a solution (x,y) of $x^2 - 5y^2 = \pm 1$. We would like to show that $(x,y) = f_+^n(2,0)$ for some nonnegative integer n, which is equivalent to $(2,0) = f_-^n(x,y)$. As usual, let n be the largest nonnegative integer such that $(x',y') \coloneqq f_-^n(x,y)$ has coordinates which are nonnegative integers. The second step establishes that $(x')^2 - 5(y')^2 = \pm 4$.

   Now, not both coordinates of $f_-(x',y')$ are nonnegative integers, so either $5y' < x'$ or $x' < y'$. On one hand, if $x' < y'$, then
$$1 = (x')^2 - 5(y')^2 < -4(y')^2 \leq 0,$$
   which makes no sense. On the other hand, if $5y' < x'$, then
$$1 = (x')^2 - 5(y')^2 > 25(y')^2 - 5(y')^2 = 20(y')^2,$$
   so $y' = 0$, so $(x',y') = (2,0)$, which completes the proof. ∎

One can now compute solutions to $x^2 - 5y^2 = \pm 1$ from Example 2.5 by checking which pairs (x_n, y_n) have both coordinates even.

Example 2.6. Define the sequence of ordered pairs of nonnegative integers $\{(x_n,y_n)\}_{n=0}^\infty$ recursively by $(x_0,y_0) \coloneqq (1,0)$ and
$$(x_{n+1}, y_{n+1}) \coloneqq (2x_n + 5y_n, x_n + 2y_n)$$
for any $n \geq 0$. Then for any pair of nonnegative integers (x,y) such that $x^2 - 5y^2 = \pm 1$, we have $(x,y) = (x_n,y_n)$ for some nonnegative integer n, and (x_n,y_n) is a solution for each n. In fact, $x_n^2 + x_n y_n - y_n^2 = (-1)^n$ for each n.

Solution. Define (x_n', y_n') to be the recursion of Example 2.5; recall that $x_n' \equiv y_n' \pmod 2$ always. We claim that (x_n', y_n') has both terms even if and only if n is divisible by 3. Indeed, applying the recursion three times, we see
$$(x,y) \mapsto \left(\frac{x+5y}{2}, \frac{x+y}{2}\right) \mapsto \left(\frac{3x+5y}{2}, \frac{x+3y}{2}\right) \mapsto (2x+5y, x+2y), \tag{2.1}$$

and $x \equiv y \equiv 2x+5y \equiv x+2y \pmod 2$. Thus, the first three terms are $(2,0)$, $(1,1)$, and $(3,1)$, so an induction shows that (x_n, y_n) has both terms even if and only if n is divisible by 3, as needed.

It follows that having $x^2 - 5y^2 = \pm 1$ must have $(x,y) = (x'_{3n}/2, y'_{3n}/2)$ for some nonnegative integer n, and we have $(x'_{3n}/2)^2 - 5(y'_{3n}/2)^2 = (-1)^n$ for each n. To complete the proof, we note that the sequence $\{(x'_{3n}/2, y'_{3n})\}_{n=0}^{\infty}$ has $(x'_0/2, y'_0/2) = (1,0)$ and recursion given by

$$\left(x'_{3(n+1)}/2, y'_{3(n+1)}/2\right) = (2x'_{3n}/2 + 5y'_{3n}/2, x'_{3n}/2 + 2y'_{3n}/2)$$

by the computation in (2.1). The result follows. ■

2.1.2 Pell Equations with Sophistication

Let's explain what's going on in the previous examples. Example 2.2 is quite mysterious because we have removed the context from which

$$f_{\pm}(x,y) = (2x \pm 3y, \pm x + 2y)$$

came from. To explain this, the key is to factor our equation $x^2 - 3y^2 = 1$ into

$$\left(x - y\sqrt{3}\right)\left(x + y\sqrt{3}\right) = 1.$$

Even though we are now working with $\sqrt{3}$s, this is good because now the problem is completely multiplicative! By a little brute force, one finds the solution $(x,y) = (2,1)$, so for example $\left(2-\sqrt{3}\right)\left(2+\sqrt{3}\right) = 1$. But now the solution $2+\sqrt{3}$ allows us to find more solutions because $x^2 - 3y^2 = 1$ implies

$$\begin{aligned} 1 &= \left(x + y\sqrt{3}\right)\left(x - y\sqrt{3}\right) \\ &= \left(x + y\sqrt{3}\right)\left(2+\sqrt{3}\right)\left(x - y\sqrt{3}\right)\left(2-\sqrt{3}\right) \\ &= \left((2x+3y) + (x+2y)\sqrt{3}\right)\left((2x+3y) - (x+2y)\sqrt{3}\right). \end{aligned}$$

We now see where f_+ came from, and f_- comes from multiplying out $\left(x + y\sqrt{3}\right)\left(2-\sqrt{3}\right)$ to produce another solution. Note that this also immediately explains why f_+ and f_- are inverse operations with no work: f_+ takes $x + y\sqrt{3}$ and multiplies by $2+\sqrt{3}$, but then f_- multiplies by $2-\sqrt{3}$, for a total multiplication by $\left(2+\sqrt{3}\right)\left(2-\sqrt{3}\right) = 1$.

One can now translate Example 2.3 into saying that all solutions (x,y) in the nonnegative integers to $x^2 - 3y^2 = 1$ take the form $x + y\sqrt{d} = \left(2+\sqrt{3}\right)^n$ for some nonnegative integer n. With this in mind, the argument of Example 2.2 directly generalizes into the following result.

Proposition 2.7. Let d be a non-square positive integer, and suppose that $x^2 - dy^2 = 1$ has a positive integer solution. Let (x_1, y_1) be the minimal positive integer solution in y. Then a pair of integers (x,y) satisfies $x^2 - dy^2 = 1$ if and only if there is a sign $\varepsilon \in \{\pm 1\}$ and an integer $n \in \mathbb{Z}$ such that

$$x + y\sqrt{d} = \varepsilon\left(x_1 + y_1\sqrt{d}\right)^n.$$

Proof. Before doing anything, we check that all the given pairs (x,y) are in fact solutions. Well, an induction on n shows that

$$x - y\sqrt{d} = \varepsilon\left(x_1 - y_1\sqrt{d}\right)^n,$$

so

$$x^2 - dy^2 = \left(x - y\sqrt{d}\right)\left(x + y\sqrt{d}\right) = \varepsilon^2\left(x_1 - y_1\sqrt{d}\right)^n\left(x_1 + y_1\sqrt{d}\right)^n = \left(x_1^2 - dy_1^2\right)^n = 1.$$

It remains to show that any (x,y) satisfying $x^2 - dy^2 = 1$ have the desired form.

We quickly reduce to the case where $x, y \geq 0$. By adjusting ε, we may assume that $x \geq 0$; it remains to show that $x + y\sqrt{d} = \left(x_1 + y_1\sqrt{d}\right)^n$ for some $n \in \mathbb{Z}$. Further, if $x^2 - dy^2 = 1$, then $\left(x - y\sqrt{d}\right)\left(x + y\sqrt{d}\right) = 1$,

so $\left(x+y\sqrt{d}\right)^{-1} = x-y\sqrt{d}$. Thus, $x+y\sqrt{d} = \left(x_1+y_1\sqrt{d}\right)^n$ if and only if $x-y\sqrt{d} = \left(x_1+y_1\sqrt{d}\right)^{-n}$, so we may assume that $y \geq 0$ as well. It remains to show $x+y\sqrt{d} = \left(x_1+y_1\sqrt{d}\right)^n$ for some $n \geq 0$.

Because (x_1, y_1) is a solution in positive integers, we see that $x_1+y_1\sqrt{d} \geq 1+\sqrt{d} > 1$, so $\left(x_1+y_1\sqrt{d}\right)^n$ is an increasing sequence as n varies over nonnegative integers. Thus, for any $x+y\sqrt{d}$, we may find some $n \geq 0$ such that

$$\left(x_1+y_1\sqrt{d}\right)^n \leq x+y\sqrt{d} < \left(x_1+y_1\sqrt{d}\right)^{n+1},$$

so

$$1 \leq \left(x+y\sqrt{d}\right)\left(x_1+y_1\sqrt{d}\right)^{-n} < x_1+y_1\sqrt{d}.$$

We would now like to compare our solutions and use minimality of (x_1, y_1) to conclude. This will require the following result.

Lemma 2.8. Let d be a non-square positive integer. Given nonnegative integer solutions (a_1, b_1) and (a_2, b_2) to $x^2-dy^2 = 1$, the following are equivalent.

(a) $a_1+b_1\sqrt{d} \geq a_2+b_2\sqrt{d}$.

(b) $a_1 \geq a_2$ and $b_1 \geq b_2$.

(c) $a_1 \geq a_2$ or $b_1 \geq b_2$.

The same statements are equivalent once $\geq$ is replaced with $>$.

Proof. Of course, (b) implies (a) and (c). Additionally, if $a^2-db^2 = 1$, then $a = \sqrt{1+db^2}$ and $b = \frac{1}{d}\sqrt{a^2-1}$, both of which are strictly increasing functions, so (c) implies (b). With this in mind, we note that $a^2-db^2 = 1$ will imply $a+b\sqrt{d} = \sqrt{1+db^2}+b\sqrt{d}$, a function strictly increasing in b, so (a) implies (c), completing the proof. ∎

Remark 2.9. In light of Lemma 2.8, a solution (x, y) to $x^2-dy^2 = 1$ in the positive integers which is minimal in y is equivalently minimal in x. (In fact, any reasonable weighting will continue to make (x, y) the smallest solution; see Problem 2.1.2.) As such, we may sloppily say that such a solution is simply "minimal" or "smallest" in the future instead of specifying what it is minimal with respect to.

Now, we define

$$x'+y'\sqrt{d} := \left(x+y\sqrt{d}\right)\left(x_1-y_1\sqrt{d}\right)^n = \left(x+y\sqrt{d}\right)\left(x_1+y_1\sqrt{d}\right)^{-n}.$$

We quickly check that $x', y' \geq 0$. The main point is that $x'+y'\sqrt{d} \geq 1$ and $(x')^2-d(y')^2 = 1$ by construction. For example, $y' = \pm\frac{1}{d}\sqrt{(x')^2-1}$, so if $x' < 0$, then $x'+y'\sqrt{d} < x'+\frac{1}{\sqrt{d}}(x') < 0$. Similarly, note that $(x')^2-d(y')^2 = 1$, so $\left(x'+y'\sqrt{d}\right)\left(x'-y'\sqrt{d}\right) = 1$, so $y' < 0$ implies that $\sqrt{d} \leq x'-y'\sqrt{d} = 1/\left(x'+y'\sqrt{d}\right) \leq 1$, which does not make sense.

Now, $x', y' \geq 0$ and the inequalities

$$1 \leq x'+y'\sqrt{d} < x_1+y_1\sqrt{d}$$

enforces $1 \leq x' < x_1$ and $0 \leq y' < y_1$ by Lemma 2.8. Because y_1 is minimal among positive integer solutions, we must have $y' = 0$, so $x' = 1$ is forced. It follows that $x'+y'\sqrt{d} = 1$, so

$$x+y\sqrt{d} = \left(x_1+y_1\sqrt{d}\right)^n$$

follows. ∎

Remark 2.10. We will show in Proposition 2.13 that the hypothesis that $x^2 - dy^2 = 1$ having a solution is always fulfilled.

Continuing with our examples, Example 2.3 is explained by factoring the equation $x^2 - 2y^2 = \pm 1$ into

$$\left(x - y\sqrt{2}\right)\left(x + y\sqrt{2}\right) = \pm 1.$$

We can now use the fact that $\left(1 + \sqrt{2}\right)\left(1 - \sqrt{2}\right) = -1$ and thus $\left(1 + \sqrt{2}\right)\left(-1 + \sqrt{2}\right) = 1$ to build the relevant $f_\pm$: we compute

$$\left(x + y\sqrt{2}\right)\left(\pm 1 + \sqrt{2}\right) = \left(\pm x + 2y, x \pm y\right),$$

which explains f_+ and f_-, and as a bonus, we still explain why f_+ and f_- are inverse functions.

What made Example 2.3 more complicated than Example 2.2 is that our "smallest solution" $1+\sqrt{2}$ did not have $\left(1 + \sqrt{2}\right)\left(1 - \sqrt{2}\right)$ equal to 1 but instead equal to -1. We could have worked with $\left(1 + \sqrt{2}\right)^2 = 3+2\sqrt{2}$ instead, but this would in some sense be dishonest: $3+2\sqrt{2}$ is not really the smallest solution to an equation of the type $x^2 - dy^2 = c$. One can reasonably ask why

$$x^2 - 3y^2 = -1$$

has no solution, a question answered by looking $\pmod 4$. In contrast, $x^2 - 2y^2 = -1$ has no such local obstruction.

Example 2.5 continues the trend of making the equation more complicated. This time, we want to factor $x^2 - 5y^2 = \pm 4$ as

$$\left(\frac{x + y\sqrt{5}}{2}\right)\left(\frac{x - y\sqrt{5}}{2}\right) = \pm 1,$$

where the point is that our "smallest solution" is given by $\left(\frac{1+\sqrt{5}}{2}\right)\left(\frac{1-\sqrt{5}}{2}\right) = -1$. The presence of these half-integers might seem disconcerting; for example, the product of two numbers of the form $\frac{a}{2} + \frac{b}{2}\sqrt{5}$ need not take that form. However, Example 2.5 finds that we are only interested in numbers of the form $\frac{a}{2} + \frac{b}{2}\sqrt{5}$ where a and b have the same parity, and we can check that

$$\left\{\frac{a}{2} + \frac{b}{2}\sqrt{5} : a, b \in \mathbb{Z} \text{ have the same parity}\right\}$$

is closed under addition and multiplication. As such, we can build $f_\pm$ as in Example 2.3: $\left(\frac{1+\sqrt{5}}{2}\right)\left(\frac{-1+\sqrt{5}}{2}\right) = 1$ means we want to compute

$$\left(\frac{x + y\sqrt{5}}{2}\right)\left(\frac{\pm 1 + \sqrt{5}}{2}\right) = \frac{\frac{\pm x + 5y}{2} + \frac{x \pm y}{2}\sqrt{5}}{2},$$

where $\frac{\pm x+5y}{2}$ and $\frac{x\pm y}{2}$ are both integers because x and y have the same parity.

There are a number of aspects of the above explanations which are still mysterious, but we will respond to them in time. For example, why did we not consider elements of the form $\frac{a}{2} + \frac{b}{2}\sqrt{3}$ in Example 2.2? For that matter, why not elements of the form $\frac{a}{3} + \frac{b}{3}\sqrt{5}$ in Example 2.5? More fundamentally we keep pulling these small solutions like $2 + \sqrt{3}$ and $1 + \sqrt{2}$ and $\frac{1+\sqrt{5}}{2}$ out from nowhere, so where do they come from? This last question is the one we will focus on answering first.

2.1.3 Using Continued Fractions

In this subsection, we will discuss how to find the smallest solution (x, y) to an equation of the form

$$x^2 - dy^2 = c$$

in some restricted cases. The hope is that one can then use this smallest solution to produce all other solutions as in Examples 2.2, 2.3 and 2.5.

Motivated by the previous section, we factor our equation into

$$\left(x - y\sqrt{d}\right)\left(x + y\sqrt{d}\right) = c.$$

Now, the point is that, if c is small, we need $x - y\sqrt{d}$ to also be abnormally small. Thus, x/y needs to be a good rational approximation of $\sqrt{d}$. If x/y is good enough, we can use Theorem 1.64 in order to deduce that x/y is a continued fraction convergent of $\sqrt{d}$. This motivates us to use continued fraction convergents. For example, continued fraction convergents automatically produce small values for $x^2 - dy^2$.

Lemma 2.11. Let $\{h_n/k_n\}_{n=0}^{\infty}$ be the sequence of continued fraction convergents of $\sqrt{d}$, where d is a non-square positive integer. Then

$$\left|h_n^2 - dk_n^2\right| < 2\sqrt{d} + 1$$

for any $n \geq 0$.

Proof. By Proposition 1.40, we see that

$$\left|\sqrt{d} - \frac{h_n}{k_n}\right| < \frac{1}{k_n^2},$$

so factoring as above yields

$$\left|h_n^2 - dk_n^2\right| = \left|h_n - k_n\sqrt{d}\right| \cdot \left|h_n + k_n\sqrt{d}\right| < \frac{\left|h_n + k_n\sqrt{d}\right|}{k_n} = \left|\sqrt{d} + \frac{h_n}{k_n}\right|.$$

To complete our bounding, we use the triangle inequality, writing

$$\left|\sqrt{d} + \frac{h_n}{k_n}\right| \leq 2\sqrt{d} + \left|\sqrt{d} - \frac{h_n}{k_n}\right| < 2\sqrt{d} + \frac{1}{k_n^2} \leq 2\sqrt{d} + 1,$$

so we are done. ∎

Remark 2.12. We could alternately recover the above bound by using the bound on $s_\bullet$ given in the proof of Proposition 1.55. The difference here is rather inconsequential.

A pigeonhole argument can use Lemma 2.11 to show that $x^2 - dy^2 = 1$ at least has solutions.

Proposition 2.13. Let d be a positive integer. Then the equation $x^2 - dy^2 = 1$ has a solution in the positive integers.

Proof. Applying the pigeonhole principle to Lemma 2.11, there exists some $N \in \mathbb{Z}$ with $|N| < 2\sqrt{d} + 1$ such that there are infinitely many positive integer solutions (x, y) to $x^2 - dy^2 = N$. Note that there are no positive integer solutions to $x^2 - dy^2 = 0$, so $N \neq 0$. We would like to pick up two solutions (x_1, y_1) and (x_2, y_2) to $x^2 - dy^2 = N$ and write

$$\frac{x_1 + y_1\sqrt{d}}{x_2 + y_2\sqrt{d}}$$

to produce an element with $x^2 - dy^2 = 1$. However, the above element need not take the form $a + b\sqrt{d}$ where a and b are positive integers. Thus, for technical reasons, we note that there are only finitely many elements in $(\mathbb{Z}/|N|\mathbb{Z})^2$, so we must be able to find two distinct pairs of positive integers (x_1, y_1) and (x_2, y_2) such that $x_1 - dy_1^2 = x_2^2 - dy_2^2 = N$ and $x_1 \equiv x_2 \pmod{N}$ and $y_1 \equiv y_2 \pmod{N}$. (Roughly speaking, this $\pmod{N}$ business is necessary because of Problem 2.1.1.)

We will now be able to divide $x_1 + y_1\sqrt{d}$ by $x_2 + y_2\sqrt{d}$. To see this, note there exist integers a and b such that

$$x_1 \pm y_1\sqrt{d} = x_2 \pm y_2\sqrt{d} + N\left(a \pm b\sqrt{d}\right).$$

Thus,

$$\frac{x_1 \pm y_1\sqrt{d}}{x_2 \pm y_2\sqrt{d}} = 1 + \frac{N}{x_2 \pm y_2\sqrt{d}} \cdot \left(a \pm b\sqrt{d}\right) = 1 + \left(x_2 \mp y_2\sqrt{d}\right)\left(a \pm b\sqrt{d}\right).$$

The right-hand side here can thus be expressed as $x \pm y\sqrt{d}$ where x and y are some integers, from which we find

$$\left(x + y\sqrt{d}\right)\left(x - y\sqrt{d}\right) = \left(\frac{x_1 + y_1\sqrt{d}}{x_2 - y_2\sqrt{d}}\right)\left(\frac{x_1 + y_1\sqrt{d}}{x_2 - y_2\sqrt{d}}\right) = \frac{N}{N} = 1.$$

It remains to check that we can coerce (x, y) to live in the positive integers while remaining solutions to $x^2 - dy^2 = 1$. Note that $x = 0$ would imply that $0^2 - dy^2 = 1$, which is impossible. Similarly, $y = 0$ would imply that $x = \pm 1$ and so $(x_2, y_2) = \pm(x_1, y_1)$, which cannot be the case because these are distinct pairs of positive integers. Now, with $x, y \neq 0$, we note that we may adjust their signs to assume that $x, y > 0$ while remaining a solution to $x^2 - dy^2 = 1$, completing the proof. ∎

Even though the argument of Proposition 2.13 is somewhat obnoxious, now that we know we have some solution, we can find fairly efficiently using continued fraction convergents, finally executing the approach given at the start of this subsection.

Proposition 2.14. Let d be a non-square positive integer, and suppose (x, y) is a pair of positive integers such that $\left|x^2 - dy^2\right| < \sqrt{d}$. Then x/y is a continued fraction convergent of $\sqrt{d}$.

Proof. We have two cases.

- Suppose $x^2 - dy^2 > 0$. The main point is the bounding

$$\left|\sqrt{d} - \frac{x}{y}\right| = \frac{\left|\frac{x^2}{y^2} - d\right|}{\left|\sqrt{d} + \frac{x}{y}\right|} = \frac{\left|x^2 - dy^2\right|}{\left|\sqrt{d} + \frac{x}{y}\right|} \cdot \frac{1}{y^2} < \frac{\sqrt{d}}{\left|\sqrt{d} + \frac{x}{y}\right|} \cdot \frac{1}{y^2}.$$

  Now, $x^2 - dy^2 > 0$ implies $x/y > \sqrt{d}$, so we have an upper-bound of $\frac{\sqrt{d}}{2\sqrt{d}} \cdot \frac{1}{y^2} < \frac{1}{2y^2}$, allowing us to conclude by Theorem 1.64.

- Suppose $x^2 - dy^2 < 0$. In this case, we will actually show that y/x is a continued fraction convergent of $1/\sqrt{d}$, which will be enough: if $\sqrt{d} = [a_0; a_1, a_2, \ldots]$, then $1/\sqrt{d} = [0; a_0, a_1, a_2, \ldots]$, so the reciprocal of a nonzero continued fraction convergent of $1/\sqrt{d}$ will be a continued convergent of $\sqrt{d}$. Anyway, we bound

$$\left|\frac{1}{\sqrt{d}} - \frac{y}{x}\right| = \left|\frac{x - y\sqrt{d}}{x\sqrt{d}}\right| = \frac{\left|x^2 - dy^2\right|}{x\sqrt{d} \cdot \left|x + y\sqrt{d}\right|} < \frac{1}{\left|1 + y/x\sqrt{d}\right|} \cdot \frac{1}{x^2}.$$

  Now, $x^2 - dy^2 < 0$ implies $y/x > 1/\sqrt{d}$, so $\left|1 + y/x\sqrt{d}\right| > 2$, so our error is at most $\frac{1}{2x^2}$, allowing us to conclude by Theorem 1.64. ∎

Remark 2.15. For example, to find solutions to $x^2 - dy^2 = 1$, we see that we must look among continued fraction convergents $\{h_n/k_n\}_{n=0}^{\infty}$ of $\sqrt{d}$, and the periodicity of Remark 1.57 assures us that we might as well check within $0 \leq n \leq 4d$ or so.

Let's use Proposition 2.14 for fun and profit.

Example 2.16. A pair of integers (x, y) satisfies $x^2 - 19y^2 = 1$ if and only if there is a sign $\varepsilon \in \{\pm 1\}$ and an integer $n \in \mathbb{Z}$ such that

$$x + y\sqrt{d} = \varepsilon \left(170 + 39\sqrt{19}\right)^n.$$

Solution. In light of Proposition 2.7, it suffices to show that the smallest solution (x, y) to $x^2 - dy^2 = 1$ is $170 + 39\sqrt{19}$. By Proposition 2.14, it suffices to examine the continued fraction convergents of $\sqrt{19}$ for solutions to $x^2 - dy^2 = 1$. To compute this continued fraction, we use Proposition 1.55. This produces the (large) table as follows.

n	-2	-1	0	1	2	3	4	5	6	7
r_n			0	4	2	3	3	2	4	4
s_n			1	3	5	2	5	3	1	3
a_n			4	2	1	3	1	2	8	$\cdots$
h_n	0	1	4	9	13	48	61	170	$\cdots$	$\cdots$
k_n	1	0	1	2	3	11	14	39	$\cdots$	$\cdots$

Because $(r_7, s_7) = (r_1, s_1)$, we see that $\sqrt{19} = [4; \overline{2, 1, 3, 1, 2, 8}]$ by the proof of Corollary 1.56. Anyway, we recall from the proof of Proposition 1.55 that $s_n = (-1)^n \left(h_{n-1}^2 - dk_{n-1}^2\right)$, so we are looking for $s_n = 1$, for which we see the first nontrivial solution happens at $s_6 = 1$, so $p_5^2 - 19q_5^2 = 1$ is our smallest solution. Referencing the table, this is $(x, y) = (170, 39)$, as needed. ■

Example 2.17. The equation $x^2 - 34y^2 = -1$ has no integer solutions.

Solution. By Proposition 2.14, it suffices to check continued fraction convergents $\{h_n/k_n\}_{n=0}^{\infty}$ of $\sqrt{34}$, so we use Proposition 1.55 to compute the continued fraction of $\sqrt{34}$, producing the following table.

n	0	1	2	3	4	5
r_n	0	5	4	4	5	5
s_n	1	9	2	9	1	9
a_n	5	1	4	1	10	$\cdots$

Now, $(r_1, s_1) = (r_5, s_5)$ implies that $(r_{n+4}, s_{n+4}) = (r_n, s_n)$ for any $n \geq 1$ by the recurrence in Proposition 1.55. Because $s_n = (-1)^n \left(h_{n-1}^2 - 34k_{n-1}^2\right) > 0$ by the proof of Proposition 1.55 and Corollary 1.56, we see that any solution $h_n^2 - 194k_n^2 = -1$ must have n even while $s_{n+1} = 1$. However, the above periodicity shows that $s_n = 1$ if and only if $n \equiv 0 \pmod 4$, so n will never be odd. So there are no solutions to $x^2 - 34y^2 = -1$. ■

Remark 2.18. We remark that $3^2 - 34 \cdot 1^2 = -5^2$ and $5^2 - 34 \cdot 1^2 = -3^2$, so $(3/5, 1/5)$ and $(5/3, 1/3)$ are both rational solutions to $x^2 - 34y^2 = -1$. We thus claim that $x^2 - 34y^2 \equiv -1 \pmod n$ has a solution for each positive integer n, thus breaking a strong form of the local-to-global principle. Indeed, using the Chinese remainder theorem, choose integers (x, y) such that

$$(x, y) \equiv \begin{cases} (3/5, 1/5) \pmod{p^{\nu_p(n)}} & \text{if } p \neq 5, \\ (5/3, 1/3) \pmod{p^{\nu_p(n)}} & \text{if } p = 5. \end{cases}$$

Note that this is in fact finitely many modular conditions because $p^{\nu_p(n)} = 1$ for all primes p except for the p such that $p \mid n$. Anyway, the construction yields $x^2 - 34y^2 \equiv -1 \pmod{p^{\nu_p(n)}}$ for all primes p, which yields $x^2 - 34y^2 \equiv -1 \pmod n$ by the Chinese remainder theorem.

2.1.4 Generalized Pell Equations

Proposition 2.7 combined with the method of Proposition 2.14 tells us how to solve equations of the form

$$x^2 - dy^2 = 1$$

where d is a non-square positive integer. We now use these solutions to solve

$$x^2 - dy^2 = c$$

in general. When $|c| < \sqrt{d}$, we can still use Proposition 2.14, but for larger c, this method does not work. There is still a method to use continued fractions (not of $\sqrt{d}$ but of a similar quadratic irrational) in order to look for solutions, but because we have not discussed an efficient algorithm to compute such continued fractions, we will settle for the following effective result. The proof technique we use below will be used again in various levels of generality.

Proposition 2.19. Fix a non-square positive integer d, and let (x_0, y_0) be a pair of positive integers such that $x_0^2 - dy_0^2 = 1$. Set $u \coloneqq x_0 + y_0\sqrt{d}$. For any nonzero integer c, any integral solution (x, y) of $x^2 - dy^2 = c$ can be written as $x + y\sqrt{d} = (x' + y'\sqrt{d})u^n$ for some integer n where

$$|x'| \le \frac{\sqrt{|c|}\,(\sqrt{u}+1)}{2} \qquad \text{and} \qquad |y'| \le \frac{\sqrt{|c|}\,(\sqrt{u}+1)}{2\sqrt{d}}.$$

Proof. Roughly speaking, we would like to measure the height of a nonzero quadratic irrational of the form $x + y\sqrt{d}$. Additionally, we would like for multiplication of the quadratic irrationals to add heights together (to make our arithmetic easier), and we would like for our heights to be positive. It would make sense to set our height, then, to be $\log\left|x + y\sqrt{d}\right|$, but from an algebraic point of view, we would like to put $\sqrt{d}$ and $-\sqrt{d}$ on equal footing. As such, we define

$$\operatorname{Log}(x + y\sqrt{d}) \coloneqq \left(\log\left|x + y\sqrt{d}\right|, \log\left|x - y\sqrt{d}\right|\right) \in \mathbb{R}^2,$$

where $x, y \in \mathbb{Q}$ are not both zero. (Note that the value of $x + y\sqrt{d}$ determines x and y because $\sqrt{d}$ is irrational.) Here are some basic properties of Log.

- For any $(x, y), (x', y') \in \mathbb{Q}^2 \setminus \{(0, 0)\}$, a direct expansion yields

$$\begin{aligned}\left(x + y\sqrt{d}\right)\left(x' + y'\sqrt{d}\right) &= (xx' + dyy') + (xy' + yx')\sqrt{d}\\ \left(x - y\sqrt{d}\right)\left(x' - y'\sqrt{d}\right) &= (xx' + dyy') - (xy' + yx')\sqrt{d}.\end{aligned}$$

  Thus,

$$\begin{aligned}\operatorname{Log}\left((x + y\sqrt{d})(x' + y'\sqrt{d})\right) &= \left(\log\left|(x + y\sqrt{d})(x' + y'\sqrt{d})\right|, \log\left|(x - y\sqrt{d})(x' - y'\sqrt{d})\right|\right)\\ &= \left(\log\left|x + y\sqrt{d}\right|, \log\left|x - y\sqrt{d}\right|\right) + \left(\log\left|x' + y'\sqrt{d}\right|, \log\left|x' - y'\sqrt{d}\right|\right)\\ &= \operatorname{Log}(x + y\sqrt{d}) + \operatorname{Log}(x - y\sqrt{d}).\end{aligned}$$

  For example, an induction implies that $\operatorname{Log}\left((x + y\sqrt{d})^n\right) = n\operatorname{Log}(x + y\sqrt{d})$ for any $(x, y) \in \mathbb{Q}^2 \setminus \{(0, 0)\}$.

- For any (x, y) such that $x^2 - dy^2 = 1$, we see that $(x + y\sqrt{d})(x - y\sqrt{d})$, so

$$\log\left|x + y\sqrt{d}\right| + \log\left|x - y\sqrt{d}\right| = 0.$$

Thus, $\operatorname{Log}\left(u^k\right) \subseteq \left\{(x,y) \in \mathbb{R}^2 : x+y=0\right\}$. In other words, the vectors $\operatorname{Log}(u)$ and $(1,1)$ are orthogonal.

More generally, if $a^2 - db^2 = c$ for any nonzero integer c, then $H(a+b\sqrt{d})$ will output to the plane $\left\{(x,y) \in \mathbb{R}^2 : x+y = \log|c|\right\}$, which is the set of vectors whose dot product with $(1,1)$ is $\log|c|$.

Now, suppose that $x^2 - dy^2 = c$, and for brevity let $\alpha \coloneqq x + y\sqrt{d}$ and $\overline{\alpha} \coloneqq x - y\sqrt{d}$ so that $\alpha\overline{\alpha} = c$. The point is to estimate the needed n in the proposition by pushing everything through Log. Note $(x,y) \neq (0,0)$ because c is nonzero, so we may place $\operatorname{Log}(\alpha) \in \mathbb{R}^2$. The second point above tells us that $\operatorname{Log}(u)$ and $(1,-1)$ form an orthogonal basis of $\mathbb{R}^2$, so we write

$$\operatorname{Log}(\alpha) = s\operatorname{Log}(u) + t(1,1),$$

which is

$$(\log|\alpha|, \log|\overline{\alpha}|) = (s\log u + t, -s\log u + t).$$

As roughly explained in the second point above, we can quickly solve for t: summing coordinates, we see that $2t = \log|\alpha\overline{\alpha}|$, so $t = \frac{1}{2}\log|c|$.

We now estimate n as the closest integer to s so that $|n-s| \leq \frac{1}{2}$. This allows us to define $x' + y'\sqrt{d} \coloneqq \alpha u^{-n}$, and we see that x' and y' are integers because $\alpha = x + y\sqrt{d}$ and $u^{-1} = x_0 - y_0\sqrt{d}$ have all coefficients integers. For brevity, define $\alpha' \coloneqq x' + y'\sqrt{d}$ and $\overline{\alpha'} \coloneqq x' - y'\sqrt{d}$. Note $x + y\sqrt{d} = \alpha u^n$ by construction. It remains to prove the inequalities on x' and y'. The idea is to bound α' and $\overline{\alpha'}$ first by passing through Log, we see

$$\left(\log|\alpha'|, \log\left|\overline{\alpha'}\right|\right) = \operatorname{Log}(\alpha') = \operatorname{Log}(\alpha) - n\operatorname{Log}(u) = \left((s-n)\log u + \frac{1}{2}\log|c|, -(s-n)\log u + \frac{1}{2}\log|c|\right).$$

One of $(s-n)$ or $-(s-n)$ is nonnegative. In the case $s-n \geq 0$, then we see that $|\alpha'| = u^{s-n}|c|^{1/2} \leq \sqrt{u|c|}$, and $\left|\overline{\alpha'}\right| = u^{-s-n}|c|^{1/2} \leq \sqrt{|c|}$ because $u > 1$. A symmetric pair of inequalities hold in the case where $s-n \leq 0$, so in total we find

$$|\alpha'| + \left|\overline{\alpha'}\right| \leq \sqrt{|c|}\left(\sqrt{u}+1\right). \tag{2.2}$$

We are now ready to bound x' and y'. Well, note

$$|x'| = \left|\frac{\alpha' + \overline{\alpha'}}{2}\right| \leq \frac{|\alpha'| + \left|\overline{\alpha'}\right|}{2},$$

and

$$|y'| = \left|\frac{\alpha' - \overline{\alpha'}}{2\sqrt{d}}\right| \leq \frac{|\alpha'| + \left|\overline{\alpha'}\right|}{2\sqrt{d}},$$

from which (2.2) completes the argument. ∎

Example 2.20. A pair of integers (x,y) satisfies $x^2 - 19y^2 = 5$ if and only if there are signs $\varepsilon_x, \varepsilon_y \in \{\pm 1\}$ and an integer $n \in \mathbb{Z}$ such that

$$x + y\sqrt{d} = \left(\varepsilon_x 9 + \varepsilon_y 2\sqrt{19}\right)\left(170 + 39\sqrt{19}\right)^n.$$

Solution. We feed $u \coloneqq 170 + 39\sqrt{19}$ into Proposition 2.19; recall we found u in Example 2.16. By Proposition 2.19, we would like to find solutions to $x^2 - 19y^2 = 5$ where

$$|y| \leq \frac{\sqrt{5}\left(\sqrt{170 + 39\sqrt{19}} + 1\right)}{2\sqrt{19}}.$$

We claim that the right-hand side is less than 6; a more precise calculation could show that it is less than 5, but this will make little difference to us. Squaring it is equivalent to show that

$$\left(\sqrt{170+39\sqrt{19}}+1\right)^2 < \frac{4\cdot 19\cdot 6^2}{5}.$$

Well,

$$\left(\sqrt{170+39\sqrt{19}}+1\right)^2 < \left(\sqrt{170+40\cdot 5}+1\right)^2 < \left(\sqrt{400}+1\right)^2 = 441 < 456 = 4\cdot 19\cdot 6,$$

which is good enough. We now deal with $|y|<6$ via the following table.

y	$5+19y^2$	$x=\pm\sqrt{5+19y^2}$
0	5	not integer
1	24	not integer
2	81	± 9
3	176	not integer
4	309	not integer
5	480	not integer

Thus, $(x,y)=(\varepsilon_x 9,\varepsilon_y 2)$ are only solutions with $|y|<6$. The result follows from Proposition 2.19. ■

Example 2.21. There are no integer solutions (x,y) to $x^2-19y^2=-5$.

Solution. We work as in Example 2.20. Because $|5|=|-5|$, the entire argument goes through until the table. Here is our new table.

y	$-5+19y^2$	$x=\pm\sqrt{-5+19y^2}$
0	-5	not integer
1	14	not integer
2	71	not integer
3	156	not integer
4	299	not integer
5	470	not integer

Thus, there are no solutions (x,y) to $x^2-19y^2=-5$ in the region $|y|<6$, which is enough to complete the proof by Proposition 2.19 and the computations of Example 2.20. ■

2.1.5 A Harder Problem

Here is a statement of the same form as Example 2.2.

Proposition 2.22. Define the sequence of ordered triples of nonnegative integers $\{(x_0,y_0,z_0)\}_{n\in\mathbb{Z}}$ recursively by $(x_0,y_0,z_0):=(1,0,0)$ and

$$(x_{n+1},y_{n+1},z_{n+1})=(x_n+2y_n+2z_n,x_n+y_n+2z_n,x_n+y_n+z_n)$$

for any $n\in\mathbb{Z}$. Then for any triple (x,y,z) of integers, we have $x^3+2y^3+4z^3-6xyz=1$ if and only if $(x,y,z)=(x_n,y_n,z_n)$ for integer n.

We do not yet have the machinery to show this result, though it does look like an elementary argument similar to the ones we have already given ought to suffice. A primary goal of our next few weeks is to be able to figure out where statements like Proposition 2.22 come from and provide some idea how to solve them. This will take us through a little algebraic number theory.

2.1.6 Problems

Do ten points worth of the following exercises. You must do the survey question at the end.

Problem 2.1.1 (2 points). Show that there exist a positive integer d and pairs of positive integers (x_1, y_1) and (x_2, y_2) such that $x_1^2 - dy_1^2 = x_2^2 - dy_2^2$ even though the ratio

$$\frac{x_1 + y_1\sqrt{d}}{x_2 + y_2\sqrt{d}}$$

does not take the form $a + b\sqrt{d}$ where $a, b \in \mathbb{Z}$.

Problem 2.1.2 (3 points). Let d be a non-square positive integer, and fix weights $\alpha, \beta \in \mathbb{R}_{\geq 0}$. Given nonnegative integer solutions (a_1, b_1) and (a_2, b_2) to $x^2 - dy^2 = 1$, show that the following are equivalent.

(a) $\alpha a_1 + \beta b_1 \geq \alpha a_2 + \beta b_2$.

(b) $a_1 \geq a_2$.

Problem 2.1.3 (4 points). Define the sequence of ordered pairs of nonnegative integers $\{(x_n, y_n)\}_{n=0}^{\infty}$ recursively by $(x_0, y_0) := (1, 0)$ and

$$(x_{n+1}, y_{n+1}) := (y_n, x_n + y_n)$$

for any $n \geq 0$. Then for any pair of nonnegative integers (x, y) such that $x^2 + xy - y^2 = \pm 1$, show that $(x, y) = (x_n, y_n)$ for some nonnegative integer n.

Problem 2.1.4 (4 points). We study the equation $x^2 - 223y^2 = -3$.

(a) Show that the equation $x^2 - 223y^2 = -3$ has no integer solutions.

(b) Show that the equation $x^2 - 223y^2 = -27$ does have integer solutions. Conclude that, for any integer n coprime to 3, the equation $x^2 - 223y^2 \equiv -3 \pmod{n}$ has solutions.

The quickest way to dispatch of the prime 3 is via Problem 4.1.1.

Problem 2.1.5 (5 points). We study the equation $x^2 - 61y^2 = 3$.

(a) Describe all positive integer solutions (x, y) to $x^2 - 61y^2 = 3$.

(b) Using (a), compute the three smallest positive integers y such that $61y^2 + 3$ is a perfect square. (A little care is required.)

You will almost certainly need to use a computer program; if you use a computer program, please submit it.

Problem 2.1.6 (7 points). Fix a non-square positive integer d, and let $\sqrt{d} = [a_0; a_1, a_2, \ldots]$ be the continued fraction with $\{h_n/k_n\}_{n=0}^{\infty}$ as the continued fraction convergents. Let $m \geq 2$ be the least positive integer such that $h_{m-1}^2 - dk_{m-1}^2 = 1$, which exists by Proposition 2.14.

(a) Using the notation of Proposition 1.55, show that $s_m = 1$ and thus $r_{m+1} = a_0$ and $s_{m+1} = d - a_0^2$.

(b) Show that $(a_{n+m}, r_{n+m}, s_{n+m}) = (a_n, r_n, s_n)$ for all $n \geq 1$.

(c) Show that (x, y) is a positive integer solution to $x^2 - dy^2 = 1$ if and only if $(x, y) = (h_{nm-1}, k_{nm-1})$ for some $n \geq 1$.

Problem 2.1.7 (0 points). Please rate the speed of the following lectures, from "much too slow" to "much too fast."

- October 2: Integrality, I
- October 4: Integrality, II
- October 6: Discriminants

Please also rate the difficulty of the problems on the homework you solved.

2.2 Number Rings

In our solutions to Pell equations, we frequently ran into numbers of the form

$$a + b\sqrt{d}$$

where $a, b \in \mathbb{Z}$ and d is a positive integer which is not a square. But in our explanation of Example 2.5 we even ran into numbers of the form

$$\frac{a + b\sqrt{5}}{2}$$

where a and b had the same parity. The goal of this section is to contextualize what is going on here. Starting in this section, we will assume basic ring theory, on the level of any reasonable abstract algebra text. We refer to Appendix A.2 for the necessary field theory.

2.2.1 Normal Domains

It is a question of classical interest in number theory to take an integral domain and ask if it is a unique factorization domain: in some sense, unique factorization domains are "the best" rings (e.g., they are computationally nice to work with because their multiplicative structure can be well-understood). However, it is in general somewhat difficult to do such a check, and one spends a good part of an algebraic number theory learning how to do so.

To start off, a basic hypothesis is that the integral domain be "normal." For motivation, we recall the Rational root theorem.

Theorem 2.23 (rational root for $\mathbb{Z}$). Let $f(x) \in \mathbb{Z}[x]$ be a monic polynomial with integer coefficients. If q is a rational root of $f(x)$, then q is an integer.

More generally, we will prove the following more general result.

Theorem 2.24 (rational root). Let A be a unique factorization domain with fraction field K, and let $f(x) \in A[x]$ be a monic polynomial with coefficients in A. If $q \in K$ is a root of $f(x)$, then $q \in A$.

Proof. Quickly, if $q = 0$, there is nothing to say. Otherwise, by using unique factorization, we may write $q = a/b$ where a and b have no irreducible factors in common. Explicitly, by unique factorization, we may express q as a quotient of nonzero elements in A by writing

$$q = \frac{u \prod_{i=1}^{n} p_i^{\alpha_i}}{\prod_{j=1}^{n} p_i^{\beta_i}},$$

where u is a unit, $\alpha_\bullet$ and $\beta_\bullet$ are nonnegative integers, and each $p_\bullet$ is a unique irreducible (not equal to the product of any other $p_\bullet$ and a unit). Then we may write

$$q = u \underbrace{\prod_{\substack{1 \le i \le n \\ \alpha_i > \beta_i}} p^{\alpha_i - \beta_i}}_{a:=} \Big/ \underbrace{\prod_{\substack{1 \le i \le n \\ \alpha_i < \beta_i}} p^{\beta_i - \alpha_i}}_{b:=}$$

Notably, if $\alpha_i = \beta_i$, then we may remove p_i.

We now proceed with the usual proof of the Rational root theorem. Write

$$f(x) = x^d + \sum_{k=0}^{d-1} r_k x^k$$

where $d = \deg f$ and $r_0, r_1, \dots, r_{d-1} \in A$. We are given that $f(a/b) = 0$. As such, the main point is that we can manipulate $f(a/b) = 0$ into

$$0 = a^d + \sum_{k=0}^{d-1} r_k a^k b^{d-k} = a^d + b \sum_{k=0}^{d-1} r_k a^k b^{(d-1)-k}.$$

To show that $q \in A$, we will show that b is a unit, which indeed implies that $q = ab^{-1} \in A$. Because A is a unique factorization domain, it suffices to show that no irreducible element p divides b. Well, if $p \mid b$, then p divides the sum above, so reducing $\pmod p$ requires $p \mid a^d$, and so $p \mid a$ because p is prime by Lemma A.15; however, no irreducible divides both a and b by their construction, so we are done. ■

Let's see how Theorem 2.24 can detect when a ring is fails to be a unique factorization domain.

Example 2.25. Construct the ring $\mathbb{Z}[\sqrt{5}] := \{a + b\sqrt{5} : a, b \in \mathbb{Z}\}$ where addition and multiplication are as expected. Then $\mathbb{Z}[\sqrt{5}]$ is not a unique factorization domain.

Solution. Quickly, we check that $\mathbb{Z}[\sqrt{5}]$ is in fact a subring of (say) $\mathbb{C}$: we have all the needed identities, and we are closed under the needed operations because

$$\left(a + b\sqrt{5}\right) + \left(a' + b'\sqrt{5}\right) = (a + a') + (b + b')\sqrt{5},$$
$$\left(a + b\sqrt{5}\right) \cdot \left(a' + b'\sqrt{5}\right) = (aa' + 5bb') + (ab' + ba')\sqrt{5}.$$

(We will not check that $\mathbb{Z}[\alpha_1, \dots, \alpha_n]$ is a ring in the future.) Anyway, the main content of this example is to consider the polynomial

$$f(x) := x^2 - x - 1,$$

which is monic and has integer coefficients (in particular, the coefficients are in $\mathbb{Z}[\sqrt{5}]$). Using the quadratic formula, we see that $\frac{1+\sqrt{5}}{2}$ is a root of $f(x)$ which lives in the quotient field of $\mathbb{Z}[\sqrt{5}]$ but not in $\mathbb{Z}[\sqrt{5}]$. Thus, Theorem 2.24 tells us that $\mathbb{Z}[\sqrt{5}]$ cannot be a unique factorization domain! ■

Exercise 2.26. More generally, let $d \equiv 1 \pmod 4$ be an integer which is not a square. Then show that the set

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

is a subring of $\mathbb{C}$ but fails to be a unique factorization domain because $\frac{1+\sqrt{d}}{2}$ is the root of a monic polynomial with integer coefficients.

The test we are applying is worth turning into an adjective.

Definition 2.27 (normal)**.** Fix an integral domain A with fraction field K. Then A is said to be *normal* if and only if the following holds: for any monic polynomial $f(x) \in A[x]$, if $q \in K$ is a root of $f(x)$, then $q \in A$.

Example 2.28. Theorem 2.24 is equivalent to the statement that unique factorization domains are normal.

Non-Example 2.29. The content of Example 2.25 is showing that $\mathbb{Z}[\sqrt{5}]$ is not normal.

It does turn out that the rings $\mathbb{Z}[\sqrt{2}]$ and $\mathbb{Z}[\sqrt{3}]$ are normal, but we will hold off showing this until we have built a little more theory.[1]

2.2.2 Number Rings

Notably, we showed that $\mathbb{Z}[\sqrt{5}]$ is not normal by only looking at monic polynomials with coefficients in $\mathbb{Z}$. This is surprising because the definition of a normal domain allows our coefficients to live in the full ring $\mathbb{Z}[\sqrt{5}]$, but we only used $\mathbb{Z}$! It will turn out that using $\mathbb{Z}$ is enough, though showing this requires a little effort. Regardless, we are motivated to make the following definitions.

Definition 2.30 (algebraic integer)**.** Fix a field extension K of $\mathbb{Q}$. An element $\alpha \in K$ is an *algebraic integer* if and only if α is the root of some monic polynomial in $\mathbb{Z}[x]$.

Definition 2.31 (number ring)**.** Fix a finite field extension K of $\mathbb{Q}$. Then the *number ring* $\mathcal{O}_K$ of K consists of all the algebraic integers in K.

Example 2.32. We see that $\mathcal{O}_{\mathbb{Q}} = \mathbb{Z}$ by Theorem 2.24. Explicitly, any element $n \in \mathbb{Z}$ is the root of the monic polynomial $x - n \in \mathbb{Z}[x]$. On the other hand, if $\alpha \in \mathbb{Q}$ is an algebraic integer, then α is the root of a monic polynomial in $\mathbb{Z}[x]$, so $\alpha \in \mathbb{Z}$ by Theorem 2.24.

Though we will not dwell on it too much, it is worth acknowledging that the following generalized (relative) notion is more correct to work with.

Definition 2.33 (integral)**.** Fix an embedding of rings $A \subseteq B$. An element $\alpha \in B$ is *integral over* A if and only if α is the root of some monic polynomial in $A[x]$.

Example 2.34. Fix a field extension K of $\mathbb{Q}$. Then $\alpha \in K$ is an algebraic integer if and only if α is integral over $\mathbb{Z}$.

[1] For example, it is possible to show that these rings are unique factorization domains directly. This approach does not generalize to further $\mathbb{Z}[\sqrt{d}]$ because, for example, $\mathbb{Z}[\sqrt{15}]$ is normal but not a unique factorization domain.

We do need to check that $\mathcal{O}_K$ is in fact a ring. Of course, 0 and 1 are algebraic integers (they are the roots of the monic polynomials x and $x-1$, respectively), so it remains to show that $\mathcal{O}_K$ is closed under addition and multiplication. This requires a little work. The following result is known as the "determinant trick" in commutative algebra. Note that this result (and its corollaries) is basically Proposition A.29.

Proposition 2.35. Fix an embedding of rings $A \subseteq B$ and some $\alpha \in B$. Then the following are equivalent.

(a) α is integral over A.

(b) The ring $A[\alpha]$ is finitely generated as an A-module.

(c) There is a subring $A' \subseteq B$ finitely generated as an A-module containing both A and α.

Here, a ring A' containing A is finitely generated as an A-module (or "finitely generated over A") if and only if there are finitely many elements $r_1', \dots, r_n'$ such that each $r' \in A'$ can be written as

$$r' = \sum_{k=1}^{n} a_k r_k'$$

for some $a_1, \dots, a_n \in A$. In other words, each element of A' is an A-linear combination of some fixed finite set in A'.

Proof. We show the implications separately, following Proposition A.29.

- To see that (a) implies (b), we suppose α is the root of the monic polynomial $f(x) \in A[x]$ of degree d, written as
$$f(x) = x^d + \sum_{k=0}^{d-1} a_k x^k.$$
Then, for any $n \geq d$, we can express α^n as a $\mathbb{Z}$-linear combination of lower powers because
$$\alpha^n = -\sum_{k=0}^{d-1} a_k \alpha^{k+d-n}.$$
It follows that $A[\alpha]$ is generated by the elements $1, \alpha, \alpha^2, \dots, \alpha^{d-1}$.
- Note that (b) implies (c) by setting $A' = A[\alpha]$.
- Checking that (c) implies (a) is harder. If $A' = 0$, then $\alpha = 0$, so there is nothing to say; otherwise, $A' \neq 0$. Suppose A' is generated by the elements $r_1', r_2', \dots, r_n'$. Note that $\alpha r_i' \in A'$ for each r_i', so we may write
$$\alpha r_i' = \sum_{j=1}^{n} a_{ij} r_j'$$
for some elements $a_{ij} \in A$. In other words, the matrix $T \coloneqq (a_{ij})_{i,j=1}^n$ has
$$\alpha \begin{bmatrix} r_1' \\ \vdots \\ r_n' \end{bmatrix} = T \begin{bmatrix} r_1' \\ \vdots \\ r_n' \end{bmatrix}.$$
Thus, $T - \alpha I_n$ is an $n \times n$ matrix with entries in A', and it has the nonzero vector $(r_1', \dots, r_n')$ in its kernel, so $\det(T - \alpha I_n) = 0$. Expanding out the polynomial $\det(\alpha I_n - T) = 0$ makes α the root of a monic polynomial (of degree n) with coefficients in A, so α is indeed integral over A. ∎

In our application, we will also want the following lemma.

Lemma 2.36. Let $A \subseteq B \subseteq C$ be embeddings of rings. If C is finitely generated as a B-module, and B is finitely generated as an A-module, then C is finitely generated as an A-module.

Proof. The point is to concatenate our generating sets together; indeed, this argument is basically the same as Lemma A.20. Say that B is generated over A by the elements $b_1, \dots, b_m$, and say that C is generated over B by the elements $c_1, \dots, c_n$. Then any $c \in C$ has some $b_1', \dots, b_n' \in B$ such that

$$c = \sum_{\ell=1}^{n} b_\ell' c_\ell,$$

but now each b_ℓ' can be expanded over A as

$$c = \sum_{\ell=1}^{n} \sum_{k=1}^{m} a_{k\ell} b_k c_\ell$$

for some $a_{k\ell} \in A$. Thus, the elements $b_k c_\ell$ for $1 \le k \le m$ and $1 \le \ell \le n$ generate C over A, so we are done. ∎

Corollary 2.37. Fix a finite field extension K of $\mathbb{Q}$. Then $\mathcal{O}_K$ is a normal ring.

Proof. This argument is essentially Corollary A.30. We run our checks separately.

- We check that $\mathcal{O}_K$ is a ring. As discussed previously, $0, 1 \in \mathcal{O}_K$ because these elements are the roots of the polynomials x and $x - 1$, respectively. It remains to show that, for any $\alpha, \beta \in \mathcal{O}_K$, we have $\alpha + \beta, \alpha\beta \in \mathcal{O}_K$. The main point is to show that $\mathbb{Z}[\alpha, \beta]$ is finitely generated as a $\mathbb{Z}$-module, which will complete the proof by Proposition 2.35 because $\alpha + \beta, \alpha\beta \in \mathbb{Z}[\alpha, \beta]$.

  Well, let α and β be the roots of the monic polynomials $f(x), g(x) \in \mathbb{Z}[x]$ respectively. Then by Proposition 2.35 shows that $\mathbb{Z}[\beta]$ is finitely generated as a $\mathbb{Z}$-module, and $f(\alpha) = 0$ shows that α is integral over $\mathbb{Z}[\beta]$, so $\mathbb{Z}[\alpha, \beta]$ is finitely generated as a $\mathbb{Z}[\beta]$-module. We conclude $\mathbb{Z}[\alpha, \beta]$ is finitely generated as a $\mathbb{Z}$-module by Lemma 2.36.

- We check that $\mathcal{O}_K$ is normal. Suppose that $\alpha \in K$ is the root of the monic polynomial $f(x) \in \mathcal{O}_K[x]$; we show that $\alpha \in \mathcal{O}_K$ by showing that α is an algebraic integer. Well, expand $f(x)$ as

  $$f(x) = x^d + \sum_{k=0}^{d-1} a_k x^k$$

  for some $a_0, \dots, a_{d-1} \in \mathcal{O}_K$. Each $a_\bullet$ is integral over $\mathbb{Z}$, so Proposition 2.35 tells us that $\mathbb{Z}[a_\bullet]$ for each $a_\bullet$. As such, as in the previous check, we may build the tower

  $$\mathbb{Z} \subseteq \mathbb{Z}[a_0] \subseteq \mathbb{Z}[a_0, a_1] \subseteq \cdots \subseteq \mathbb{Z}[a_0, \dots, a_{d-1}],$$

  where each ring is finitely generated over the previous one by Proposition 2.35. Then Lemma 2.36 tells us that $\mathbb{Z}[a_0, \dots, a_{d-1}]$ is finitely generated as a $\mathbb{Z}$-module. Lastly, $f(\alpha) = 0$ tells us that α is integral over $\mathbb{Z}[a_0, \dots, a_{d-1}]$, so $\mathbb{Z}[a_0, \dots, a_{d-1}, \alpha]$ is finitely generated as a $\mathbb{Z}[a_0, \dots, a_{d-1}]$-module and hence finitely generated as a $\mathbb{Z}$-module by Lemma 2.36, meaning that α is integral over $\mathbb{Z}$ by Proposition 2.35. ∎

As another application of Proposition 2.35, we dispose of the following annoying technicality: as discussed in Lemma A.27, any algebraic element in a field extension will be the root of some unique irreducible polynomial. When integral, we would like this polynomial to have integral coefficients, but this is not technically obvious, so we place this check into the following lemma.

Lemma 2.38. Fix a finite field extension K of $\mathbb{Q}$ and some $\alpha \in \mathcal{O}_K$. Let $f(x) \in \mathbb{Q}[x]$ be the unique monic irreducible polynomial with $f(\alpha) = 0$. Then $f(x) \in \mathbb{Z}[x]$.

Proof. Embed K into $\mathbb{C}$, via (say) Proposition A.31. Then $f(x)$ will factor in $\mathbb{C}$ as

$$f(x) = \prod_{i=1}^{n} (x - \alpha_i).$$

Now, we know that $\alpha \in \mathcal{O}_K$ is the root of some monic $g(x) \in \mathbb{Z}[x]$, so Lemma A.27 tells us that $g(x) = f(x)q(x)$ for some $q(x) \in \mathbb{Q}[x]$; comparing leading coefficients, we see that the leading coefficient of q is also 1. Now, for any root $\alpha_i \in \mathbb{C}$ of f, we see that $g(\alpha_i) = 0$, so $\alpha_i \in \mathbb{C}$ is also an algebraic integer.

Thus, we consider the ring $A \coloneqq \mathbb{Z}[\alpha_1, \ldots, \alpha_n]$. Iteratively applying Proposition 2.35, the fact that each α_i is integral implies that A is finitely generated as a $\mathbb{Z}$-module, and so every element of A is integral over $\mathbb{Z}$. But direct expansion shows that the coefficients of f live in A and hence are all integral over $\mathbb{Z}$ and hence are integers by Example 2.32. ■

Remark 2.39. It follows from Proposition A.34 and Corollary A.35 that the trace and norm of algebraic integer is an integer. Indeed, the stated results show that these values arise as integer multiples of coefficients of the minimal polynomial, and the above lemma shows that the minimal polynomial has integer coefficients.

Example 2.40. Fix $K \coloneqq \mathbb{Q}(\sqrt[3]{2})$. Then

$$\mathrm{N}_{K/\mathbb{Q}}\left(a + b\sqrt[3]{2} + c\sqrt[3]{4}\right) = a^3 + 2b^3 + 4c^3 - 6abc.$$

In particular, if $a, b, c \in \mathbb{Z}$, then this is an integer.

Solution. One way to solve this is via Corollary A.35 and direct expansion. A slightly more conceptual way to do this is by directly appealing to the definition of the norm. Set $\alpha \coloneqq a + b\sqrt[3]{2} + c\sqrt[3]{4}$ for brevity. Then note that the elements $1, \sqrt[3]{2}, \sqrt[3]{4}$ form a basis for K as a $\mathbb{Q}$-vector space (see, for example, Lemma A.28), so the computations

$$\begin{aligned}
\alpha &= a + b\sqrt[3]{2} + c\sqrt[3]{4}, \\
\alpha\sqrt[3]{2} &= 2c + a\sqrt[3]{2} + b\sqrt[3]{4}, \\
\alpha\sqrt[3]{4} &= 2b + 2c\sqrt[3]{2} + a\sqrt[3]{4},
\end{aligned}$$

tell us that we can represent the mutliplication-by-α map $K \to K$ by the matrix

$$\begin{bmatrix} a & 2c & 2b \\ b & a & 2c \\ c & b & a \end{bmatrix}.$$

One can directly compute that the determinant of this matrix is $a\left(a^2 - 2bc\right) - b\left(2ac - 2b^2\right) + c\left(4c^2 - 2ab\right)$, which simplifies correctly. ■

2.2.3 Number Rings of Quadratic Extensions

After doing all that theory, we are owed an example, so we will compute $\mathcal{O}_K$ for quadratic field extensions $K = \mathbb{Q}(\sqrt{d})$ of $\mathbb{Q}$. We will want the following lemma, which we have been using in some guise for quite a bit of the previous section.

Lemma 2.41. Fix $K := \mathbb{Q}(\sqrt{d})$ where d is a non-square integer. Then the function $\sigma\colon K \to K$ given by $\sigma\left(a+b\sqrt{d}\right) := a - b\sqrt{d}$ is a ring homomorphism.

Proof. To begin, note that σ is well-defined because $a+b\sqrt{d} = a'+b'\sqrt{d}$ implies that $(a-a') = (b'-b)\sqrt{d}$ and so $a=a'$ and $b=b'$ because $\sqrt{d}$ is irrational. To check that σ is a homomorphism, we note that $\sigma(0)=0$ and $\sigma(1)=1$ and

$$\begin{aligned}\sigma\left((a+b\sqrt{d})+(a'+b'\sqrt{d})\right) &= (a+a')-(b+b')\sqrt{d}\\ &= \sigma(a+b\sqrt{d})+\sigma(a'+b'\sqrt{d})\\ \sigma\left((a+b\sqrt{d})(a'+b'\sqrt{d})\right) &= (aa'+dbb')-(ab'+ba')\sqrt{d}\\ &= \sigma\left(a+b\sqrt{d}\right)\sigma\left(a'+b'\sqrt{d}\right),\end{aligned}$$

completing the proof. ∎

Example 2.42. Fix $K := \mathbb{Q}(\sqrt{2})$. Then $\mathbb{Z}[\sqrt{2}] = \mathcal{O}_K$. In particular, $\mathbb{Z}[\sqrt{2}]$ is normal.

Proof. Note that $\sqrt{2} \in \mathcal{O}_K$ because it is the root of the polynomial $x^2-2=0$. Thus, $\mathbb{Z}[\sqrt{2}]$ is finitely generated as a $\mathbb{Z}$-module by Proposition 2.35, and we see $\mathbb{Z}[\sqrt{2}] \subseteq \mathcal{O}_K$.

We now show that $\mathcal{O}_K \subseteq \mathbb{Z}[\sqrt{2}]$, which is harder. Suppose $a+b\sqrt{2} \in \mathcal{O}_K$, where we allow $a,b \in \mathbb{Q}$. We want to show that $a,b\in\mathbb{Z}$. Well, $a+b\sqrt{2}$ is the root of some monic polynomial $f(x)\in\mathbb{Z}[x]$, so $\sigma\left(a+b\sqrt{2}\right) = a-b\sqrt{2}$ is also the root of $f(x)$ by Lemma 2.41, so $a-b\sqrt{2}$ is also an algebraic integer. Thus,

$$\begin{aligned}(a+b\sqrt{2})+(a-b\sqrt{2}) &= 2a,\\ \left((a+b\sqrt{2})-(a-b\sqrt{2})\right)\sqrt{2} &= 4b,\\ (a+b\sqrt{2})(a-b\sqrt{2}) &= a^2-2b^2,\end{aligned}$$

are also algebraic integers. But they are also rational and hence actually integers by Example 2.32, so we may write $a=a_0/2$ and $b=b_0/4$ for some integers a_0 and b_0. Then

$$a^2-2b^2 = \frac{a_0^2}{4}-\frac{b_0^2}{8} = \frac{2a_0^2-b_0^2}{8}$$

needs to be an integer, so $2a_0^2 \equiv b_0^2 \pmod 8$, which can only happen if $a_0 \equiv 0 \pmod 2$ and $b_0 \equiv 0 \pmod 4$. Thus, a and b are in fact integers, so we conclude. ∎

More generally, we can show the following statement.

Proposition 2.43. Fix a non-square squarefree integer d, and fix $K := \mathbb{Q}(\sqrt{d})$. Then we have

$$\mathcal{O}_K = \begin{cases}\mathbb{Z}[\sqrt{d}] & \text{if } d \equiv 2,3 \pmod 4,\\ \mathbb{Z}\left[\frac{1+\sqrt{d}}{2}\right] & \text{if } d \equiv 1 \pmod 4.\end{cases}$$

Proof. We proceed as in the example. Of course, $\sqrt{d}\in\mathcal{O}_K$ because it is the root of $x^2-d=0$, and if $d\equiv 1 \pmod 4$, then $\frac{1+\sqrt{d}}{2} \in \mathcal{O}_K$ because it is the root of

$$x^2 - x - \frac{d-1}{4} = 0.$$

Thus, Proposition 2.35 then assures us that $\mathbb{Z}[\sqrt{d}] \subseteq \mathcal{O}_K$ and $\mathbb{Z}\left[\frac{1+\sqrt{d}}{2}\right] \subseteq \mathcal{O}_K$ when $d \equiv 1 \pmod 4$.

It remains to show our other inclusion. Well, fix some $a + b\sqrt{d} \in \mathcal{O}_K$ where $a, b \in \mathbb{Q}$. Because $a + b\sqrt{d}$ is the root of some monic polynomial with integer coefficients, we see that $a - b\sqrt{d}$ is as well by Lemma 2.41, so $a - b\sqrt{d}$. Thus,

$$\begin{aligned}(a + b\sqrt{d}) + (a - b\sqrt{d}) &= 2a,\\ \left((a + b\sqrt{d}) - (a - b\sqrt{d})\right)\sqrt{d} &= 2bd,\\ (a + b\sqrt{d})(a - b\sqrt{d}) &= a^2 - db^2,\end{aligned}$$

are also algebraic integers. But they are also rational and hence actually integers by Example 2.32, so we may write $a = a_0/2$ and $b = b_0/(2d)$ for some integers a_0 and b_0. For example, we see that

$$4\left(a^2 - db^2\right) = a_0^2 - \frac{b_0^2}{d}$$

must be an integer, so because d is squarefree, we conclude that $d \mid b_0$, so we write $b = b_1/2$ for some integer b_1. To continue the argument, we split into cases.

- Suppose $d \equiv 2, 3 \pmod 4$. Then we see

  $$a^2 - db^2 = \frac{a_0^2 - db_1^2}{4}$$

  must be an integer. By checking $\pmod 4$, we see that both a_0 and b_1 must be even, so a and b are both integers, so $a + b\sqrt{d} \in \mathbb{Z}[\sqrt{d}]$.

- Suppose $d \equiv 1 \pmod 4$. Then we see

  $$a^2 - db^2 = \frac{a_0^2 - db_1^2}{4}$$

  must be an integer. By checking $\pmod 4$, we see that both a_0 and b_0 must have the same parity, so

  $$a + b\sqrt{d} = \frac{a_0 - b_0}{2} + a_0 \cdot \frac{1 + \sqrt{d}}{2} \in \mathbb{Z}\left[\frac{1 + \sqrt{d}}{2}\right],$$

  establishing the needed inclusion. ■

Note that Proposition 2.43 fulfills a curiosity of Example 2.5, namely about where the denominator of 2 came from and why it was so controlled!

To continue our journey of generalization to compute other rings of integers, we need to generalize aspects of the above proof. For example, the σ arises because $a + b\sqrt{d} \mapsto a - b\sqrt{d}$ is the nontrivial field embedding $\mathbb{Q}(\sqrt{d}) \hookrightarrow \mathbb{C}$ promised by Proposition A.31. Further, $2a$ and $a^2 - db^2$ are the trace and norm from appendix A.2.5, which we knew had to be integers by Remark 2.39. However, it is not so clear where the denominator of $2d$ so quickly or why was it so annoying to argue beyond that.

2.2.4 The Discriminant

The discriminant is an invariant which roughly speaking measures the size of a number field.

Definition 2.44 (discriminant). Fix a number field K of degree n over $\mathbb{Q}$, and let $\sigma_1, \dots, \sigma_n \colon K \hookrightarrow \mathbb{C}$ denote the n embeddings of Proposition A.31. Given $\alpha_1, \dots, \alpha_n \in K$, we define the *discriminant* to be

$$\operatorname{disc}(\alpha_1, \dots, \alpha_n) \coloneqq \det \begin{bmatrix} \sigma_1(\alpha_1) & \cdots & \sigma_1(\alpha_n) \\ \vdots & \ddots & \vdots \\ \sigma_n(\alpha_1) & \cdots & \sigma_n(\alpha_n) \end{bmatrix}^2.$$

Example 2.45. Let $K \coloneqq \mathbb{Q}(\sqrt{d})$ for some squarefree d. Then $\sigma_1, \sigma_2 \colon K \hookrightarrow \mathbb{C}$ are given by $a + b\sqrt{d} \mapsto a + b\sqrt{d}$ and $a + b\sqrt{d} \mapsto a - b\sqrt{d}$. Then

$$\operatorname{disc}(1, \sqrt{d}) = \det \begin{bmatrix} 1 & \sqrt{d} \\ 1 & -\sqrt{d} \end{bmatrix} = (-2\sqrt{d})^2 = 4d.$$

As stated, it is somewhat annoying to compute the discriminant or to check that it is nonzero. We will spend some time explaining how to compute this and some of its basic properties. To begin, we check that the discriminant lands in $\mathbb{Q}$ and in $\mathbb{Z}$ when the α_i live in $\mathcal{O}_K$.

Lemma 2.46. Fix a number field K of degree n over $\mathbb{Q}$. Given $\alpha_1, \dots, \alpha_n \in K$, we have

$$\operatorname{disc}(\alpha_1, \dots, \alpha_n) = \det \begin{bmatrix} \mathrm{T}_{K/\mathbb{Q}}(\alpha_1\alpha_1) & \cdots & \mathrm{T}_{K/\mathbb{Q}}(\alpha_1\alpha_n) \\ \vdots & \ddots & \vdots \\ \mathrm{T}_{K/\mathbb{Q}}(\alpha_n\alpha_1) & \cdots & \mathrm{T}_{K/\mathbb{Q}}(\alpha_n\alpha_n) \end{bmatrix}.$$

Proof. Let $\sigma_1, \dots, \sigma_n \colon K \hookrightarrow \mathbb{C}$ denote the n embeddings of Proposition A.31, and set $a_{ij} \coloneqq \sigma_i(\alpha_j)$ and $A \coloneqq (a_{ij})_{i,j=1}^n$ so that $\operatorname{disc}(\alpha_1, \dots, \alpha_n) = \det A^2$. Now, $\det A = \det A^\intercal$, so we define $B \coloneqq A^\intercal A$ so that

$$B_{ik} = \sum_{j=1}^n A_{ij}^\intercal A_{jk} = \sum_{j=1}^n \sigma_j(\alpha_i)\sigma_j(\alpha_k) = \mathrm{T}_{K/\mathbb{Q}}(\alpha_i\alpha_k).$$

Thus, the result follows because $\operatorname{disc}(\alpha_1, \dots, \alpha_n) = \det B$. ∎

Corollary 2.47. Fix a number field K of degree n over $\mathbb{Q}$ and $\alpha_1, \dots, \alpha_n \in K$. Then $\operatorname{disc}(\alpha_1, \dots, \alpha_n) \in \mathbb{Q}$. If $\alpha_1, \dots, \alpha_n \in \mathcal{O}_K$, then $\operatorname{disc}(\alpha_1, \dots, \alpha_n) \in \mathbb{Z}$.

Proof. We use Lemma 2.46. By definition, we see that $\mathrm{T}_{K/\mathbb{Q}}(\alpha_i\alpha_j) \in \mathbb{Q}$ for each i and j, so the first claim follows. If $\alpha_i \in \mathcal{O}_K$ for each i, then $\mathrm{T}_{K/\mathbb{Q}}(\alpha_i\alpha_j) \in \mathbb{Z}$ for each i and j by Remark 2.39, establishing the second claim. ∎

Proposition 2.48. Fix a number field K of degree n over $\mathbb{Q}$. Given $\alpha_1, \dots, \alpha_n \in K$, then $\operatorname{disc}(\alpha_1, \dots, \alpha_n) \neq 0$ if and only if $\alpha_1, \dots, \alpha_n$ are $\mathbb{Q}$-linearly independent.

Proof. If the α_i have a $\mathbb{Q}$-linear relation $a_1\alpha_1 + \cdots + a_n\alpha_n$, then in fact

$$a_1 \begin{bmatrix} \sigma_1(\alpha_1) \\ \vdots \\ \sigma_n(\alpha_1) \end{bmatrix} + \cdots + a_n \begin{bmatrix} \sigma_1(\alpha_n) \\ \vdots \\ \sigma_n(\alpha_n) \end{bmatrix} = 0,$$

so the rows of the matrix defining disc are linearly dependent, implying that $\operatorname{disc}(\alpha_1, \dots, \alpha_n) = 0$.

Conversely, suppose that the $\alpha_1, \dots, \alpha_n$ are $\mathbb{Q}$-linearly independent and hence form a basis of K as a $\mathbb{Q}$-vector space. Supposing for contradiction that $\operatorname{disc}(\alpha_1, \dots, \alpha_n) = 0$, then Lemma 2.46 provides a nonzero vector $(a_1, \dots, a_n) \in \mathbb{Q}^n$ such that

$$\begin{bmatrix} \mathrm{T}_{K/\mathbb{Q}}(\alpha_1\alpha_1) & \cdots & \mathrm{T}_{K/\mathbb{Q}}(\alpha_1\alpha_n) \\ \vdots & \ddots & \vdots \\ \mathrm{T}_{K/\mathbb{Q}}(\alpha_n\alpha_1) & \cdots & \mathrm{T}_{K/\mathbb{Q}}(\alpha_n\alpha_n) \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = 0,$$

so setting $\alpha \coloneqq a_1\alpha_1 + \cdots + a_n\alpha_n \neq 0$, additivity of $\mathrm{T}_{K/\mathbb{Q}}$ implies $\mathrm{T}_{K/\mathbb{Q}}(\alpha_i\alpha) = 0$ for each i. Because the α_i provide a basis for $K/\mathbb{Q}$, we conclude $\mathrm{T}_{K/\mathbb{Q}}(\beta\alpha) = 0$ for all $\beta \in K$, but this is a contradiction: set $\beta = \alpha^{-1}$ so that $\mathrm{T}_{K/\mathbb{Q}}(1) = \sigma_1(1) + \cdots + \sigma_n(1) = n$ must vanish, which is impossible. ∎

Remark 2.49. Let's explain Proposition 2.48 at a slightly higher level. Approximately speaking, the above proposition is making use of the fact that the "trace pairing" $(\alpha, \beta) \mapsto \mathrm{T}_{K/\mathbb{Q}}(\alpha\beta)$ is non-degenerate. More generally, such a statement is true when the relevant field extension is separable, which is true here because all the relevant fields are all characteristic 0—this explains why we have to do a check like $\mathrm{T}_{K/\mathbb{Q}}(1) \neq 0$ at the end. Notably, the relevant result is harder to check in positive characteristic.

So the discriminant can detect a basis of $K/\mathbb{Q}$. In fact, it can detect change of basis as well.

Lemma 2.50. Fix a number field K of degree n over $\mathbb{Q}$. Given $\mathbb{Q}$-linearly independent sets $\{\alpha_1, \dots, \alpha_n\}$ and $\{\beta_1, \dots, \beta_n\}$, let $A \in \mathbb{Q}^{n\times n}$ be the change of basis matrix with $(\alpha_1, \dots, \alpha_n) = A(\beta_1, \dots, \beta_n)$ for each i. Then

$$\operatorname{disc}(\alpha_1, \dots, \alpha_n) = (\det A)^2 \operatorname{disc}(\beta_1, \dots, \beta_n).$$

Proof. Let $\sigma_1, \dots, \sigma_n \colon K \hookrightarrow \mathbb{C}$ denote the embeddings of Proposition A.31. Because A has coordinates in $\mathbb{Q}$, we see that

$$\begin{bmatrix} \sigma_i(\alpha_1) \\ \vdots \\ \sigma_i(\alpha_n) \end{bmatrix} = A \begin{bmatrix} \sigma_i(\beta_1) \\ \vdots \\ \sigma_i(\beta_n) \end{bmatrix}$$

for each i because σ_i fixes the coordinates of A. Thus, we produce the matrix equation

$$\begin{bmatrix} \sigma_1(\alpha_1) & \cdots & \sigma_n(\alpha_1) \\ \vdots & \ddots & \vdots \\ \sigma_1(\alpha_n) & \cdots & \sigma_n(\alpha_n) \end{bmatrix} = A \begin{bmatrix} \sigma_1(\beta_1) & \cdots & \sigma_n(\beta_1) \\ \vdots & \ddots & \vdots \\ \sigma_1(\beta_n) & \cdots & \sigma_n(\beta_n) \end{bmatrix}.$$

Taking determinants and squaring completes the proof. ∎

2.2.5 Number Ring Structure

We use what we know about the discriminant to talk about number rings. The following will be our main result.

Theorem 2.51. Fix a number field K. Then $\mathcal{O}_K$ is a free abelian group of rank $[K : \mathbb{Q}]$.

We will show Theorem 2.51 by showing that $\mathcal{O}_K$ contains and is contained in a free abelian group of rank $[K : \mathbb{Q}]$. This will complete the proof by the following result.

Lemma 2.52. Fix a nonnegative integer n, and let G be a subgroup of $\mathbb{Z}^n$. Then G is a free abelian group of rank at most n.

Proof. We induct on n. If $n = 0$, there is nothing to say. If $n = 1$, then we note that any subgroup $G \subseteq \mathbb{Z}$ is closed under $\mathbb{Z}$-linear combination and is therefore an ideal and so principal by Proposition A.6, which makes $G = d\mathbb{Z}$ for some $d \in \mathbb{Z}$. Now, if $G \cong 0\mathbb{Z}$, then G is free of rank 0; and if $d \neq 0$, then $G \cong \mathbb{Z}$ is free of rank 1.

For our induction, suppose $n \geq 1$, and let $\pi \colon \mathbb{Z}^n \to \mathbb{Z}$ be the projection onto the last coordinate, and we note $\ker \pi = \mathbb{Z}^{n-1} \times \{0\}$ is isomorphic to $\mathbb{Z}^{n-1}$. Now, the main claim is that

$$G \cong (G \cap \ker \pi) \oplus \pi(G). \tag{2.3}$$

This will complete the proof because $G \cap \ker \pi \subseteq \ker \pi \cong \mathbb{Z}^{n-1}$ must be a free abelian group of rank at most $n-1$, and $\pi(G) \subseteq \mathbb{Z}$ must be a free abelian group of rank 1.

It remains to show (2.3). If $\pi(G) = \{0\}$, then $G \subseteq \ker \pi$, so $G = G \cap \ker \pi$, and there is nothing to say. Otherwise, $\pi(G) \subseteq \mathbb{Z}$ contains a nonzero element and is isomorphic to $\mathbb{Z}$, so let $h_0 \in G$ be such that $\pi(h_0) \in$

$\pi(G)$ generates $\pi(G)$. We now show $G \cong (G \cap \ker \pi) \oplus \mathbb{Z}$: define the homomorphism $\varphi\colon (G \cap \ker \pi) \oplus \mathbb{Z} \to G$ by

$$\varphi(h, n) \coloneqq h + nh_0.$$

We won't bother checking that φ is a homomorphism, but we will check that it is a bijection, which will complete the proof.

- Injective: we show trivial kernel. If $\varphi(h, n) = 0$, then $h + nh_0 = 0$, so pushing through π shows that $n\pi(h_0) = 0$, so $n = 0$. But then $h = 0$ follows.
- Surjective: for any $h \in G$, set $n \coloneqq \pi(h)/\pi(h_0)$, which is an integer by hypothesis on h_0. Then

$$\varphi(h - nh_0, n) = h - nh_0 + nh_0 = h,$$

so $h \in \operatorname{im} \varphi$. ■

Corollary 2.53. Let G be a group which both is contained in and contains a free abelian group of rank n. Then G is a free abelian group of rank n.

Proof. Write $G \subseteq \mathbb{Z}^n$. By Lemma 2.52, G is a free group of some rank $r \leq n$, so we can provide G with a basis $v_1, \dots, v_r$. Further, G contains a subgroup isomorphic to $\mathbb{Z}^n$, so let $w_1, \dots, w_n$ be the image of the basis vectors of $\mathbb{Z}^n$ in G.

We will appeal to some linear algebra to complete the proof. Embed G into $\mathbb{Q}^n$, and let V be the span of the elements of G. Notably, this is the span of the elements $v_1, \dots, v_r$. Additionally, V contains the $\mathbb{Z}$-lienarly independent elements $w_1, \dots, w_n$, but being $\mathbb{Z}$-lienarly independent implies that they must be $\mathbb{Q}$-linearly independent by clearing denominators on any relation. Thus, $r \geq \dim V \geq n$, so we are done. ■

Now, showing one direction, it is not so bad to show that $\mathcal{O}_K$ contains a free group of rank n. The point is to take any basis of $K/\mathbb{Q}$ and scale the basis vectors using the following lemma.

Lemma 2.54. Fix a number field K. For any $\alpha \in K$, there exists some positive integer n such that $n\alpha \in \mathcal{O}_K$.

Proof. By Lemma A.26, we see that α is at least algebraic over K, so we can find some monic polynomial

$$f(x) = x^d + a_{d-1}x^{d-1} + \cdots + a_1 x + a_0$$

such that $f(\alpha) = 0$, where $a_0, a_1, \dots, a_{d-1} \in \mathbb{Q}$.

The idea is to clear the denominators of f. For each i, we can write $a_i = p_i/q_i$ where q_i is a positive integer, and define $n \coloneqq q_0 q_1 \cdots q_{d-1}$. Then we see that

$$\begin{aligned} 0 &= n^d \left(\alpha^d + a_{d-1}\alpha^{d-1} + a_{d-2}\alpha^{d-2} + \cdots + a_1\alpha + a_0\right) \\ &= (n\alpha)^d + na_{d-1}(n\alpha)^{d-1} + n^2 a_{d-2}(n\alpha)^{d-2} + \cdots + n^{d-1}a_1(n\alpha) + n^d a_0, \end{aligned}$$

so we define

$$g(x) \coloneqq x^d + na_{d-1}x^{d-1} + n^2 a_{d-2} + \cdots + n^{d-1}a_1 x + n^d a_0$$

so that $g(x) \in \mathbb{Z}[x]$. Thus, $g(n\alpha) = 0$ shows that $n\alpha \in \mathcal{O}_K$. ■

Corollary 2.55. Fix a number field K. Then $\mathcal{O}_K$ contains a free abelian group of rank $[K : \mathbb{Q}]$.

Proof. For brevity, set $n \coloneqq [K : \mathbb{Q}]$, and let $\alpha_1, \dots, \alpha_n \in K$ be a basis for K as a $\mathbb{Q}$-vector space. Multiplying each α_i by a positive integer will not change the fact that this is a basis, so Lemma 2.54 allows us to assume that $\alpha_1, \dots, \alpha_n \in \mathcal{O}_K$. Now,

$$\mathbb{Z}\alpha_1 \oplus \cdots \oplus \mathbb{Z}\alpha_n \subseteq K$$

is a free abelian group of rank n (indeed, this follows because the α_i are $\mathbb{Q}$-linearly independent and hence $\mathbb{Z}$-linearly independent), and each element lives in $\mathcal{O}_K$ by Corollary 2.37. ∎

The other direction requires more effort. In particular, we now use the discriminant.

Lemma 2.56. Fix a number field K of degree n, and let $\alpha_1, \dots, \alpha_n \in \mathcal{O}_K$ generate a free abelian group of rank n as promised in Corollary 2.55. Then each $\alpha \in \mathcal{O}_K$ can be written as

$$\alpha = \frac{c_1\alpha_1 + \cdots + c_n\alpha_n}{\operatorname{disc}(\alpha_1, \dots, \alpha_n)}$$

for some $c_1, \dots, c_n \in \mathbb{Z}$.

Proof. Note that the conclusion makes sense by Proposition 2.48 because the α_i are $\mathbb{Z}$-linearly independent and hence $\mathbb{Q}$-linearly independent. So let $d \coloneqq \operatorname{disc}(\alpha_1, \dots, \alpha_n)$, which we know is a nonzero integer (see also Corollary 2.47).

The idea of the proof is to use Cramér's rule to solve for the c_i. For any $\alpha \in \mathcal{O}_K$, we do know that we can at least write

$$\alpha = q_1\alpha_1 + \cdots + q_n\alpha_n$$

for some $q_1, \dots, q_n \in \mathbb{Q}$. One equation is not enough to determine the q_i, but we can produce more: for each of the embeddings $\sigma_1, \dots, \sigma_n \colon K \hookrightarrow \mathbb{Q}$ promised by Proposition A.31, we get an equation

$$\sigma_i(\alpha) = q_1\sigma_i(\alpha_1) + \cdots + q_n\sigma_i(\alpha_n).$$

We can now use Cramér's rule or the adjugate matrix to solve for the x_i from these n different equations: we find that $q_i = a_i/b$ where $a_i, b \in \mathcal{O}_K$ and in particular $b = \det(\sigma_i(\alpha_j))_{i,j=1}^n$.

To complete the proof, we note that $b^2 = d$ by definition of d, so $q_i = ba_i/d$, so we want to set $c_i \coloneqq ba_i = q_i d$. Note $c_i = q_i d \in \mathbb{Q}$ and $c = ba_i \in \mathcal{O}_K$, so $c \in \mathcal{O}_\mathbb{Q} = \mathbb{Z}$ by Example 2.32. This completes the proof. ∎

Corollary 2.57. Fix a number field K. Then $\mathcal{O}_K$ is contained in a free abelian group of rank n.

Proof. Using the notation of Lemma 2.56, we see that

$$\mathcal{O}_K \subseteq \mathbb{Z}\frac{\alpha_1}{d} \oplus \cdots \oplus \mathbb{Z}\frac{\alpha_n}{d}$$

where $d \coloneqq \operatorname{disc}(\alpha_1, \dots, \alpha_n)$. This is what we wanted. ∎

Theorem 2.51 now follows from combining Corollaries 2.53, 2.55 and 2.57.

Remark 2.58. Note that the proof of Lemma 2.56 has essentially automated the first part of the argument of Proposition 2.43; notably, $\operatorname{disc}(1, \sqrt{d}) = 4d$ by Example 2.45, so Lemma 2.56 tells us immediately that any element of $\mathcal{O}_{\mathbb{Q}(\sqrt{d})}$ takes the form $\frac{a+b\sqrt{d}}{4d}$. Of course, more precise arguments are able to improve this.

We are now able to make the following definition.

Definition 2.59 (integral basis)**.** Fix a number field K of degree n. By Theorem 2.51, $\mathcal{O}_K$ is freely generated by n elements $\alpha_1, \dots, \alpha_n \in \mathcal{O}_K$; any such list of generators is called an *integral basis*. By abuse of notation, we will write

$$\operatorname{disc} \mathcal{O}_K \coloneqq \operatorname{disc}(\alpha_1, \dots, \alpha_n).$$

Perhaps we should check that $\operatorname{disc} \mathcal{O}_K$ does not depend on the choice of integral basis.

Lemma 2.60. Fix a number field K of degree n. Choose $\alpha_1, \dots, \alpha_n \in \mathcal{O}_K$ and $\beta_1, \dots, \beta_n \in \mathcal{O}_K$ which are $\mathbb{Q}$-linearly independent and satisfy

$$\underbrace{\mathbb{Z}\alpha_1 \oplus \cdots \oplus \mathbb{Z}\alpha_n}_{\mathcal{A}\coloneqq} \subseteq \underbrace{\mathbb{Z}\beta_1 \oplus \cdots \oplus \mathbb{Z}\beta_n}_{\mathcal{B}\coloneqq}.$$

Then there exists an integer $d = [\mathcal{B} : \mathcal{A}]$ such that $\operatorname{disc}(\alpha_1, \dots, \alpha_n) = d^2 \operatorname{disc}(\beta_1, \dots, \beta_n)$.

Proof. This follows from Lemma 2.50. Let $A \in \mathbb{Q}^{n \times n}$ be the matrix with $(\alpha_1, \dots, \alpha_n) = A(\beta_1, \dots, \beta_n)$; note this gives $|\det A| = [\mathcal{B} : \mathcal{A}]$. Because $\alpha_1, \dots, \alpha_n \in \mathbb{Z}\beta_1 \oplus \cdots \oplus \mathbb{Z}\beta_n$, we see that $A \in \mathbb{Z}^{n \times n}$. Thus, $\det A \in \mathbb{Z}$, so Lemma 2.50 finishes. ∎

In particular, if $\alpha_1, \dots, \alpha_n \in \mathcal{O}_K$ and $\beta_1, \dots, \beta_n \in \mathcal{O}_K$ are both integral bases, then their discriminants must be equal.

After all this theory, we should do another example.

Example 2.61. Fix the field $K \coloneqq \mathbb{Q}(\sqrt[3]{2})$. Then $\mathcal{O}_K = \mathbb{Z}[\sqrt[3]{2}]$, and $\operatorname{disc} \mathcal{O}_K = -108$.

Proof. This is fairly involved, so we take a deep breath. The elements $1, \sqrt[3]{2}, \sqrt[3]{4}$ are surely algebraic integers, so $\mathbb{Z}[\sqrt[3]{2}] \subseteq \mathcal{O}_K$. It remains to show the reverse inclusion. We use Lemma 2.56: to compute the discriminant, let ζ_3 denote the third root of unity, and we see

$$\begin{aligned}
\operatorname{disc}(1, \sqrt[3]{2}, \sqrt[3]{4}) &= \det \begin{bmatrix} 1 & \sqrt[3]{2} & \sqrt[3]{4} \\ 1 & \zeta_3\sqrt[3]{2} & \zeta_3^2\sqrt[3]{4} \\ 1 & \zeta_3^2\sqrt[3]{2} & \zeta_3\sqrt[3]{4} \end{bmatrix}^2 \\
&= \left(1\left(2\zeta_3^2 - 2\zeta_3^4\right) - 1\left(2\zeta_3 - 2\zeta_3^2\right) + 1\left(2\zeta_3^2 - 2\zeta_3\right)\right)^2 \\
&= 4\left(\zeta_3^2 - \zeta_3 - \zeta_3 + \zeta_3^2 + \zeta_3^2 - \zeta_3\right)^2 \\
&= 36\left(\zeta_3^2 - \zeta_3\right)^2 \\
&= -108.
\end{aligned}$$

Thus, we can write any $\alpha \in \mathcal{O}_K$ as $\frac{1}{108}(a + b\sqrt[3]{2} + \sqrt[3]{4})$ for some $a, b, c \in \mathbb{Z}$. We want to show that each of $a, b, c \in \mathbb{Z}$ is divisible by 108. There are two computations.

- If $\alpha = \frac{1}{3}(a + b\sqrt[3]{2} + \sqrt[3]{4}) \in \mathcal{O}_K$ for integers $a, b, c \in \mathbb{Z}$, we show that $3 \mid a, b, c$. By subtracting out from $1, \sqrt[3]{2}, \sqrt[3]{4}$, we may assume that $a, b, c \in \{-1, 0, 1\}$. However, by Example 2.40 tells us that the norm of α is
$$-1 < -\frac{1+2+4+6}{27} \le \frac{a^3 + 2b^3 + 4c^3 - 6abc}{27} \le \frac{1+2+4+6}{27} < 1,$$
so we must have $\mathrm{N}_{K/\mathbb{Q}}(\alpha) = 0$, which forces $\alpha = 0$ by Corollary A.35.
- If $\alpha = \frac{1}{2}(a + b\sqrt[3]{2} + \sqrt[3]{4}) \in \mathcal{O}_K$ for integers $a, b, c \in \mathbb{Z}$, we show that $2 \mid a, b, c$. This is essentially the same argument. By subtracting out from $1, \sqrt[3]{2}, \sqrt[3]{4}$, we may assume that $a, b, c \in \{0, 1\}$. However, by Example 2.40 tells us that the norm of α is
$$0 \le \frac{a^3 + 2b^3 + 4c^3 - 6abc}{8} \le \frac{1+2+4}{8} < 1,$$

so we must have $\mathrm{N}_{K/\mathbb{Q}}(\alpha) = 0$, which forces $\alpha = 0$ by Corollary A.35.

We now complete the argument. Any $\alpha \in \mathcal{O}_K$ can be written as $\alpha = \frac{1}{108}(a + b\sqrt[3]{2} + \sqrt[3]{4})$. Then $36\alpha = \frac{1}{3}(a + b\sqrt[3]{2} + \sqrt[3]{4})$, so $3 \mid a, b, c$ by the above, so we can write $\alpha = \frac{1}{36}(a' + b'\sqrt[3]{2} + c'\sqrt[3]{4})$. Considering 12α shows that $3 \mid a', b', c'$ still, and we can continue downwards until $\alpha = x + y\sqrt[3]{2} + z\sqrt[3]{4}$ for integers x, y, z. This completes the proof. ■

Remark 2.62. One could show that $\mathcal{O}_K = \mathbb{Z}[\sqrt[3]{2}]$ without using the discriminant. For example, one could compute $\operatorname{Tr} \alpha$, $\operatorname{Tr} \alpha\sqrt[3]{2}$, and $\operatorname{Tr} \alpha\sqrt[3]{4}$ to sufficiently lower-bound the denominator.

2.2.6 Problems

Do ten points worth of the following exercises.

Problem 2.2.1 (2 points)**.** Show that $\sqrt[3]{5}/2$ is not an algebraic integer.

Problem 2.2.2 (3 points)**.** We show that $\mathbb{Z}[\sqrt{-5}]$ is not a unique factorization domain.

(a) Show that the elements 2, 3, $1 + \sqrt{-5}$, and $1 - \sqrt{-5}$ are irreducible. You may find it helpful to use norms.

(b) Show that there is no unit $u \in \mathbb{Z}[\sqrt{-5}]$ such that $2u = 1 + \sqrt{-5}$ or $2u = 1 - \sqrt{-5}$.

(c) Finish the proof by noting $2 \cdot 3 = \left(1 + \sqrt{-5}\right)\left(1 - \sqrt{-5}\right)$.

Problem 2.2.3 (4 points)**.** Show that the following complex numbers α are algebraic integers by finding a monic polynomial $f(x) \in \mathbb{Z}[x]$ such that $f(\alpha) = 0$.

(a) $\alpha = \sqrt{2}$.

(b) $\alpha = \sqrt{2} + \sqrt{3}$.

(c) $\alpha = \frac{1}{3}\left(1 + \sqrt[3]{10} + \sqrt[3]{100}\right)$.

Problem 2.2.4 (4 points)**.** Let m and n be squarefree coprime integers such that $m \equiv n \equiv 1 \pmod 4$, and set $K := \mathbb{Q}(\sqrt{m}, \sqrt{n})$. Show that

$$\mathcal{O}_K = \mathbb{Z}\left[\frac{1+\sqrt{m}}{2}, \frac{1+\sqrt{n}}{2}\right].$$

Notably, $\left(\frac{1+\sqrt{m}}{2}\right)\left(\frac{1+\sqrt{n}}{2}\right) \in \mathcal{O}_K$.

Problem 2.2.5 (5 points)**.** Let $K := \mathbb{Q}(\sqrt[3]{3})$. Show that $\mathcal{O}_K = \mathbb{Z}[\sqrt[3]{3}]$, and compute $\operatorname{disc} \mathcal{O}_K$.

Problem 2.2.6 (5 points)**.** For a given positive integer n, a primitive nth root of unity is a complex number ζ such that $\zeta^n = 1$ but $\zeta^m \neq 1$ for any positive integer $m < n$.

(a) Convince yourself that $\zeta_n := e^{2\pi i/n}$ is a primitive nth root of unity.

(b) Show that ζ_n is an algebraic integer.

Problem 2.2.7 (0 points). Please rate the speed of the following lectures, from "much too slow" to "much too fast."

- October 9: Lattices
- October 11: Minkowski's Theorem
- October 13: Applications of Minkowski's Theorem

Please also rate the difficulty of the problems on the homework you solved.

2.3 Minkowski Theory

Having spent a long time building the theory of number rings, we will take a break to discuss some geometry of numbers. We will then return and prove Dirichlet's unit theorem, Theorem 2.102.

2.3.1 Lattices

The goal of the present subsection is to state and prove some basic facts about lattices in order to set us up for Minkowski's theorem. The moral of the present subsection is that lattices provide a language which allows algebra and geometry to communicate with each other.

Definition 2.63 (lattice). Fix a nonnegative integer n. Then a *lattice of rank* m Λ is a subset of $\mathbb{R}^n$ such that there exist linearly independent vectors $v_1, \dots, v_m$ with

$$\Lambda = \{a_1 v_1 + \cdots + a_m v_m : a_1, \dots, a_m \in \mathbb{Z}\}.$$

Note that $\Lambda \subseteq \mathbb{R}^n$ is a subgroup.

Example 2.64. The subset $\mathbb{Z}^2 \subseteq \mathbb{R}^2$ is a lattice. Namely, choose the linearly independent vectors $(1,0)$ and $(0,1)$, and we see that

$$\mathbb{Z}^2 = \{(a,b) : a,b \in \mathbb{Z}\} = \{a(1,0) + b(0,1) : a,b \in \mathbb{Z}\}.$$

More generally, $\mathbb{Z}^n \subseteq \mathbb{R}^n$ is a lattice for any positive integer n.

Example 2.65. The subset $\{0\} \subseteq \mathbb{R}^n$ is a lattice of rank 0 spanned by the empty set of vectors.

Remark 2.66. We take a second to remark that a lattice $\Lambda \subseteq \mathbb{R}^n$ spanned by the linearly independent vectors $v_1, \dots, v_m$ is a free abelian group of rank m. Indeed, we claim that $\varphi \colon \mathbb{Z}^m \to \Lambda$ by

$$\varphi \colon (a_1, \dots, a_m) \mapsto a_1 v_1 + \cdots + a_m v_m$$

is a group isomorphism. It is certainly a group homomorphism, and it is surjective by construction of the $v_\bullet$, so it remains to show injectivity. Well, if $\varphi(a_1, \dots, a_m) = 0$, then $a_1 v_1 + \cdots + a_m v_m = 0$, so $(a_1, \dots, a_m) = (0, \dots, 0)$ by linear independence of the $v_\bullet$.

Remark 2.67. In light of the previous remark, we will sometimes write $\Lambda = M\mathbb{Z}^n$ where $M \in \mathbb{R}^{n \times n}$ is of nonzero determinant when Λ is a lattice of rank n in $\mathbb{R}^n$.

In the sequel, we will be interested in the quotient $\mathbb{R}^n/\Lambda$, where $\Lambda \subseteq \mathbb{R}^n$ is a lattice of rank n. A convenient way to represent this is via the "fundamental parallelepiped."

Definition 2.68 (fundamental parallelepiped)**.** Fix a positive integer n and a lattice Λ of rank n in $\mathbb{R}^n$. Given a spanning set $v_1, \dots, v_n$ of Λ, the *fundamental parallelepiped* P is the set

$$\{a_1 v_1 + \cdots + a_m v_m : a_1, \dots, a_n \in [0,1)\}.$$

Example 2.69. Continue in the context of Example 2.64. Then the basis $(1,0)$ and $(0,1)$ of $\mathbb{Z}^2$ shows that

$$\{(x,y) : x, y \in [0,1)\}$$

is a fundamental parallelepiped of $\mathbb{Z}^2 \subseteq \mathbb{R}^2$.

A fundamental parallelepiped is not an invariant of the lattice Λ, but the volume is.

Lemma 2.70. Fix a positive integer n and a lattice Λ of rank n in $\mathbb{R}^n$. For any two fundamental parallelepipeds P and P' of Λ, we have $\operatorname{vol}(P) = \operatorname{vol}(P')$.

Proof. Suppose P and P' arise from the bases $v_1, \dots, v_n$ and $v_1', \dots, v_n'$, respectively, of Λ. Both of these are bases of $\mathbb{R}^n$, so there is a change of basis matrix M such that $Mv_i = v_i'$ for each i. In fact, because $v_i' \in \Lambda$, we see that the coefficients of M must be integers because all elements of Λ are $\mathbb{Z}$-linear combinations of the $v_\bullet$s. A symmetric argument provides a matrix M' with integer coefficients such that $M'v_i' = Mv_i$ for each i.

Now, the geometric interpretation of the determinant is that

$$\operatorname{vol}(P) = \left|\det \begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix}\right| \qquad \text{and} \qquad \operatorname{vol}(P') = \left|\det \begin{bmatrix} | & & | \\ v_1' & \cdots & v_n' \\ | & & | \end{bmatrix}\right|.$$

However, $\det M, \det M' \in \mathbb{Z}$ and $MM' = 1$, so $(\det M)(\det M') = 1$, so $|\det M| = 1$, so

$$\operatorname{vol}(P') = \left|\det \begin{bmatrix} | & & | \\ v_1' & \cdots & v_n' \\ | & & | \end{bmatrix}\right| = \left|\det \left(M \begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix}\right)\right| = \left|\det \begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix}\right| = \operatorname{vol}(P),$$

as desired. ∎

Remark 2.71. Intuitively, the fundamental parallelepiped P corresponding to a basis $v_1, \dots, v_n$ of a lattice $\Lambda \subseteq \mathbb{R}^n$ of rank n is the "space" taken up outside Λ. For example, any $v \in \mathbb{R}^n$ can be written uniquely as $\ell + p$ where $\ell \in \Lambda$ and $p \in P$. To see this, expand any v as

$$v = a_1 v_1 + \cdots + a_n v_n$$

where $a_1, \dots, a_n \in \mathbb{R}^n$. Then we set $\ell_i \coloneqq \lfloor a_i \rfloor$ and $p_i \coloneqq \{a_i\}$ for each i so that

$$v = a_1 v_1 + \cdots + a_n v_n = \underbrace{\ell_1 v_1 + \ldots + \ell_n v_n}_{\in \Lambda} + \underbrace{p_1 v_1 + \cdots + p_n v_n}_{\in P}.$$

To see that this expression is unique, suppose that $\ell + p = \ell' + p'$ for any $\ell, \ell' \in \Lambda$ and $p, p' \in P$. Then $p - p' = \ell' - \ell \in \Lambda$, but any vector in $p - p'$ takes the form $a_1 v_1 + \cdots + a_n v_n$ for $a_1, \dots, a_n \in (-1, 1)$ and therefore cannot be in Λ.

Anyway, the above lemma allows us to define the covolume.

Definition 2.72 (covolume)**.** Fix a positive integer n and a lattice Λ of rank n in $\mathbb{R}^n$. Then the *covolume* $\operatorname{vol}(\mathbb{R}^n/\Lambda)$ is the volume of a fundamental parallelepiped of Λ.

Example 2.73. Continue in the context of Example 2.64. Then the volume of the fundamental parallelepiped

$$\{(x,y) : x,y \in [0,1)\}$$

is the area of the square $[0,1]^2$, which is $\operatorname{vol}\left(\mathbb{R}^2/\mathbb{Z}^2\right) = 1$. More generally, $\operatorname{vol}(\mathbb{R}^n/\mathbb{Z}^n)$ is the volume of $[0,1)^n$, which is 1.

Intuitively, the covolume measures how "sparse" a lattice is. For example, in contrast to the above example, the sparser lattice $2\mathbb{Z}^2 \subseteq \mathbb{R}^2$ spanned by $\{(2,0),(0,2)\}$ has fundamental parallelepiped

$$\{(x,y) : x,y \in [0,2)\}$$

with area $4 > 1$. More generally, we are able to prove the following result.

Lemma 2.74. Fix a positive integer n and a lattice Λ of rank n in $\mathbb{R}^n$. For any invertible matrix $M \in \mathbb{R}^{n\times n}$, then

$$M\Lambda \coloneqq \{Mv : v \in \Lambda\}$$

is a lattice of rank n in $\mathbb{R}^n$ with covolume $|\det M| \operatorname{vol}(\mathbb{R}^n/\Lambda)$.

Proof. Let Λ have basis $v_1,\dots,v_n$. To check that $M\Lambda$ is a lattice of rank n, note that the vectors $Mv_1,\dots,Mv_n$ continue to be linearly independent because M is invertible, and these vectors span $M\Lambda$ because

$$\begin{aligned} M\Lambda &= \{Mv : v \in \Lambda\} \\ &= \{M(a_1v_1 + \cdots + a_nv_n) : a_1,\dots,a_n \in \mathbb{Z}\} \\ &= \{a_1Mv + \cdots + a_nMv_n : a_1,\dots,a_n \in \mathbb{Z}\}. \end{aligned}$$

It remains to compute the covolume. Well, the basis $Mv_1,\dots,Mv_n$ allows us to compute the volume of the corresponding parallelepiped, which is

$$\left|\det \begin{bmatrix} | & & | \\ Mv_1 & \cdots & Mv_n \\ | & & | \end{bmatrix}\right| = \left|\det M \cdot \det \begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix}\right| = |\det M| \operatorname{vol}(\mathbb{R}^n/\Lambda),$$

as desired. ∎

Remark 2.75. Lemma 2.74 actually tells us that $\operatorname{vol}(\mathbb{R}^n/\Lambda) > 0$. Indeed, let $v_1,\dots,v_n$ be a basis for Λ, and let M be the matrix whose columns are the $v_\bullet$; note $\det M \neq 0$ because the $v_\bullet$ are linearly independent. Now, $\Lambda = M\mathbb{Z}^n$, so $\operatorname{vol}(\mathbb{R}^n/\Lambda) = |\det M| \operatorname{vol}(\mathbb{R}^n/\mathbb{Z}^n) = |\det M|$.

It is occasionally helpful to have the following more algebraic computation of the covolume.

Lemma 2.76. Fix a positive integer n, and let $\Lambda' \subseteq \Lambda$ be free abelian groups of rank n so that $\Lambda' = M\Lambda$ for some $M \in \mathbb{Z}^{n\times n}$ of nonzero determinant. Then $|\det M| = [\Lambda : \Lambda']$.

Proof. As a starting case, we note that there is not much to say when M is diagonal. For example, if M is the diagonal matrix $(d_1,\dots,d_n)$, then

$$[\Lambda : \Lambda'] = \#\left(\frac{\mathbb{Z}^n}{d_1\mathbb{Z} \oplus \cdots \oplus d_n\mathbb{Z}}\right) = \#\left(\bigoplus_{i=1}^n \frac{\mathbb{Z}}{d_i\mathbb{Z}}\right) = |d_1 \cdots d_n| = |\det M|.$$

We now argue that applying row and column operations to M will adjust $[\Lambda : \Lambda']$ accordingly. We begin with column operations; let $v_1, \dots, v_n$ be a basis for Λ, and let $m_1, \dots, m_n$ be the columns of M so that $m_1 v_1, \dots, m_n v_n$ are a basis for Λ.

- Multiplying a column m_i of M by an integer c to create the matrix M' yields $|\det M'| = |c| \cdot |\det M|$. On the other hand, we see that $[M\Lambda' : M'\Lambda'] = |c|$, so the index is adjusted by the same factor.
- Adding a column v_i to a distinct column v_j will not change $\det M$, and it will not change Λ, so nothing happens here.

The arguments for row operations are exactly identical to the above ones, except that applying a row operation in M adjusts Λ instead of Λ'. Anyway, the point is that these row and column operations allow us to reduce to the case where M is diagonal by Gaussian elimination of matrices, which we already took care of. ■

Note the definition of a lattice provided above is quite algebraic, but we are going to get quite a bit of mileage using the geometry of lattices. To mirror this, it will be helpful to have a more geometric definition of a lattice.

Proposition 2.77. Fix a positive integer n, and let $\Lambda \subseteq \mathbb{R}^n$ be a subgroup. The following are equivalent.

(a) Λ is a lattice.

(b) There is some $R > 0$ such that $\Lambda \cap [-R, R]^n = \{0\}$.

(c) There is some $R > 0$ such that $\Lambda \cap [-R, R]^n < \infty$.

(d) For all $R > 0$, we have $\Lambda \cap [-R, R]^n < \infty$.

Proof. We show our implications in sequence.

- We show (a) implies (b). Roughly speaking, the idea is that an element in Λ with "large coefficients" must actually be "large" in $\mathbb{R}^n$. By adding linearly independent vectors to a basis of Λ, we may assume that Λ is of rank n; note that this cannot make $\Lambda \cap [-R, R]^n$ smaller for any R, so this move is safe. Now, let $v_1, \dots, v_n$ be basis for Λ, which is also a basis for $\mathbb{R}^n$, and we quickly claim that the function sending
$$a_1 v_1 + \ldots + a_n v_n \mapsto (a_1, \dots, a_n)$$
is continuous. Indeed, letting M be the matrix whose columns are the $v_\bullet$, we note that the function $(a_1, \dots, a_n) \mapsto a_1 v_1 + \cdots + a_n v_n$ is $a \mapsto Ma$, so the inverse function is $v \mapsto M^{-1} v$, which is continuous because it is linear. (Note M^{-1} exists because $\det M \neq 0$ because the $v_\bullet$ are linearly independent.)

  To continue, define the function $\|\cdot\|_\infty : \mathbb{R}^n \to \mathbb{R}$ by $\|(x_1, \dots, x_n)\|_\infty := \max\{|x_1|, \dots, |x_n|\}$, and we consider the function $f : \mathbb{R}^n \to \mathbb{R}$ given by
$$f(v) := \left\| M^{-1} v \right\|_\infty.$$
  Note that f is a continuous function (it's the $\max$ of the absolute value of some linear functions), so for $\varepsilon := 1$, there exists some $\delta > 0$ such that $\|v\| < \delta$ implies $\left\| M^{-1} v \right\|_\infty < 1$. Now, for any $v \in \mathbb{R}^n$ with $\|v\|_\infty < \delta/\sqrt{n}$, we see
$$\|v\| < \sqrt{\frac{\delta^2}{n} + \cdots + \frac{\delta^2}{n}} = \delta,$$
  so $\left| M^{-1} v \right| < 1$, so writing $v = a_1 v_1 + \ldots + a_n v_n$ must have $a_1, \dots, a_n \in (-1, 1)^n$, meaning that either $v = 0$ or $v \notin \Lambda$. In total, we see that
$$\Lambda \cap \left[-\frac{\delta}{2\sqrt{n}}, \frac{\delta}{2\sqrt{n}} \right]^n = \{0\},$$
  which completes the proof.

- Note that (b) implies (c) easily because $\{0\}$ is a finite set.
- We show (c) implies (d). We proceed by contraposition: suppose that there is some $R > 0$ such that $\Lambda \cap [-R, R]^n$ is infinite, and we will show the corresponding statement for any $r > 0$. Choose some positive integer N such that $Nr/2 > R$. Then note that

$$\bigcup_{x \in \mathbb{Z}^n \cap [-N,N]^n} \left(x + [-r/2, r/2]^n\right)$$

fully covers $[-R, R]^n$, so there must be some $x \in \mathbb{Z}^n \cap [-N, N]$ such that

$$\Lambda \cap \left(x + [-r/2, r/2]^n\right) = \infty$$

by the pigeonhole principle. Let $S \subseteq \Lambda$ denote the above infinite subset. We now translate S to the origin. Chose any fixed $v_0 \in S$, and write $v_0 = x + w_0$ where $w_0 \in [-r/2, r/2]^n$. Then for any $v \in S$, we see $v - v_0 \in \Lambda$, and writing $v = x + w$ where $w \in [-r/2, r/2]^n$ reveals that $v - v_0 = w - w_0 \in [-r, r]^n$. Thus, $S - v_0$ is an infinite subset of Λ contained in $[-r, r]^n$.

- We show (d) implies (a). Suppose that $\Lambda \subseteq \mathbb{R}^n$ is a subgroup such that $\Lambda \cap [-R, R]^n < \infty$ for all R. The main point it to find a lattice sitting inside Λ to "approximate" Λ. Let $\{v_1', \dots, v_m'\}$ be a maximal set of linearly independent vectors in Λ, and let Λ' be the lattice they span. The main claim is that $\Lambda' \subseteq \Lambda$ is a finite-index subgroup.

  To see this, let P be the fundamental parallelepiped corresponding to the basis $v_1', \dots, v_m'$. Now, Remark 2.71 tells us that any $v \in \mathbb{R}^n$ can be written uniquely as $\ell + p$ where $\ell \in \Lambda'$ and $p \in P$, so there is a function $\pi \colon \mathbb{R}^n \to P$ by sending $v = \ell + p$ to p. As such, we examine the set

$$\pi(\Lambda) \subseteq P.$$

  For each $v \in \Lambda$, we see that $v - \pi(v) \in \Lambda'$ by construction of π, so $\pi(\Lambda)$ contains a set of representatives of Λ/Λ'. On the other hand, $\pi(\Lambda) \subseteq \Lambda$ and is contained in the bounded set P, so by hypothesis on Λ, we see that $\pi(\Lambda)$ and hence Λ/Λ' is finite.

  We now complete the proof. Let $d \coloneqq [\Lambda : \Lambda']$. Then any $x \in \Lambda$ has $dx \in \Lambda'$ because Λ/Λ' is a group of order d, so $\Lambda' \subseteq \Lambda \subseteq \frac{1}{d}\Lambda'$. However, Λ' is a free abelian group of rank m, so Λ must be a free abelian group of rank m by Corollary 2.53. Let $\{v_1, \dots, v_m\}$ be a basis of Λ. These vectors span the same space that Λ' spans, which is dimension m, so we conclude that the vectors $\{v_1, \dots, v_m\}$ are linearly independent, verifying that Λ is a lattice. ■

Remark 2.78. It might be frustrating that we had to appeal to Corollary 2.53 to prove the above result. However, this is in some sense necessary because Proposition 2.77 implies Lemma 2.52: if $G \subseteq \mathbb{Z}^n$ is a subgroup, then $G \subseteq \mathbb{R}^n$ is a subgroup such that $G \cap [-1/2, 1/2]^n = \{0\}$, implying that G is a lattice in $\mathbb{R}^n$ and hence a free abelian group of rank n.

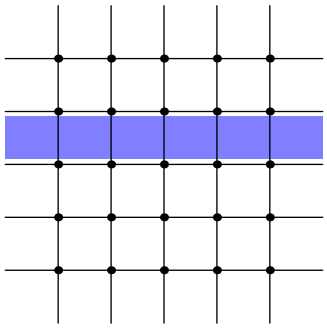
2.3.2 Minkowski's Theorem

In this subsection, we state and prove Minkowski's theorem. Our motivation comes from the following question.

Question 2.79. Let $\Lambda \subseteq \mathbb{R}^n$ be a lattice of rank n. How large must a subset $S \subseteq \mathbb{R}^n$ be to contain a lattice point in Λ?

For the time being, we will focus on the lattice $\mathbb{Z}^2 \subseteq \mathbb{R}^2$. Intuitively, if we throw a piece of Play-Doh or similar onto $\mathbb{R}^2$, we expect to hit a lattice point in $\mathbb{Z}^2$ as long as the piece of Play-Doh is large enough. We would like to rigorize this intuition.

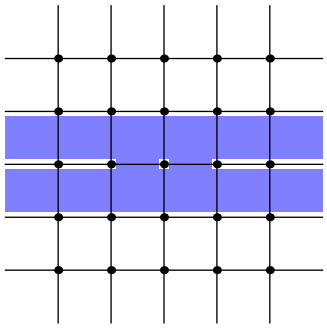
Of course, we can find subsets $S \subseteq \mathbb{R}^2$ which are very large but contain no lattice point. For example, $[0.1, 0.9] \times [-100, 100]$ has large area but no lattice point.



In order to prevent the above problem, we will require our subsets to be symmetric about the origin.

Definition 2.80 (symmetric about the origin)**.** A subset $S \subseteq \mathbb{R}^n$ is *symmetric about the origin* if and only if $x \in S$ implies $-x \in S$.

Approximately speaking, being symmetric about the origin tells us that the Play-Doh we're throwing is focused at the origin. However, we can still find subsets $S \subseteq \mathbb{R}^2$ which are very large and symmetric about the origin but contain no lattice point, as the following example shows.



The problem with the above set is that it really looks like it should contain $(0, 0)$ (as well as $(\pm 1, 0)$ and $(\pm 2, 0)$ for that matter), but we have managed to go "around" this lattice point. To remedy this, we will require our subsets to be convex.

Definition 2.81 (convex)**.** A subset $S \subseteq \mathbb{R}^n$ is convex if and only if, for any $v, w \in S$ and $t \in [0, 1]$, we have $tv + (1 - t)w \in S$. Intuitively, we are asking for the line segment connecting v and w to live in S.

We can now declare victory because being symmetric about the origin and convex does guarantee a lattice point.

Proposition 2.82. Let $\Lambda \subseteq \mathbb{R}^n$ be a lattice of rank n. Any nonempty subset $S \subseteq \mathbb{R}^n$ which is convex and symmetric about the origin contains a point of Λ.

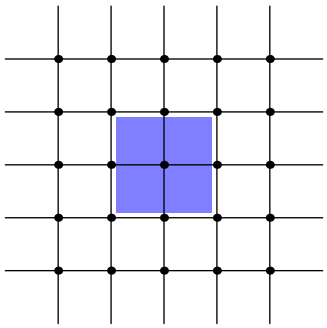
Proof. We claim that $0 \in S$. Indeed, S is nonempty, so there is some $v \in S$. But then S is symmetric about the origin, so $-v \in S$. To finish, we see that $0 = \frac{1}{2}v + \frac{1}{2}(-v)$ lives in S by convexity. ■

Notably, Proposition 2.82 is not Minkowski's theorem because this statement is quite unsatisfying: it is not fulfilling our intuition that only "large" balls of Play-Doh must hit a lattice point. Indeed, Proposition 2.82 only works because we required our Play-Doh to be focused at the origin in our symmetry condition.

As such, we have refined Question 2.79 into the following question.

Question 2.83. Let $\Lambda \subseteq \mathbb{R}^n$ be a lattice of rank n. How large must be a convex and symmetric about the origin subset $S \subseteq \mathbb{R}^n$ be in order to contain a nonzero lattice point of Λ?

Let's continue with the example $\mathbb{Z}^2 \subseteq \mathbb{R}^2$. Let $S \subseteq \mathbb{R}^2$ be convex and symmetric about the origin. The following example shows that $\text{vol}(S) \approx 4$ is permissible while still avoiding a nonzero lattice point.



The reader is welcome to try, but there isn't really a way to expand S past having $\text{vol}(S) > 4$ while avoiding a lattice point. Indeed, in some sense, the above example of $(-1,1)^2$ is a "maximal" subset of $\mathbb{R}^2$ avoiding a lattice point. Getting our constants right in arbitrary dimension, we achieve the following result.

Theorem 2.84 (Minkowski)**.** Let $\Lambda \subseteq \mathbb{R}^n$ be a lattice of rank n. Further, let $S \subseteq \mathbb{R}^n$ be convex and symmetric about the origin with

$$\text{vol}(S) > 2^n \,\text{vol}\left(\mathbb{R}^n/\Lambda\right).$$

Then S contains a nonzero lattice point in Λ.

Proof. Let $P \subseteq \mathbb{R}^n$ be some fundamental parallelepiped of Λ. The idea is to double the lattice Λ to 2Λ and consider the quotient map $\mathbb{R}^n \to \mathbb{R}^n/2\Lambda$. Concretely, one can build a basis of 2Λ by doubling a basis of Λ, so $2P$ is a fundamental parallelepiped for 2Λ. Then Remark 2.71 grants us a function $\pi\colon \mathbb{R}^n \to 2P$ by mapping $v \in \mathbb{R}^n$ to the unique $2p \in 2P$ such that $v = 2p + 2\ell$ for some $2\ell \in 2\Lambda$.

We now use the pigeonhole principle: $\text{vol}(2P) = 2^n \,\text{vol}(P) = 2^n \,\text{vol}\left(\mathbb{R}^n/\Lambda\right)$, but $\pi(S) \subseteq 2^n \,\text{vol}\left(\mathbb{R}^n/\Lambda\right)$ is compressing $\text{vol}(S)$ into a smaller volume.[2] Thus, there must be distinct $v, w \in S$ such that $\pi(v) = \pi(w)$. This is the key step of the proof. It remains to convert these vectors v and w into the desired result.

Well, $\pi(v) = \pi(w)$ implies that $v - w \in 2\Lambda$ by construction of π. Thus, there is $\ell \in \Lambda \setminus \{0\}$ such that $\ell = \frac{1}{2}v - \frac{1}{2}w$. We claim that $\ell \in S$, which will finish the proof. Well, $w \in W$ implies $-w \in S$ by being symmetric about the origin, and then $v, -w \in S$ implies

$$\ell = \frac{1}{2}v + \frac{1}{2}(-w)$$

lives in S as well by being convex. ∎

Remark 2.85. Theorem 2.84 is a really wonderful result about finding "short" vectors in a lattice, and we have seen that the result is essentially sharp. However, the result fails to actually explain how to find the nonzero vector promised. In general, one uses lattice reduction (which we will discuss a special case of in section 3.1.4) to find short vectors, but such algorithms are frequently unable to achieve the bound of Theorem 2.84.

Remark 2.86. The above proof of Theorem 2.84 technically only needs that S is "dyadic convex," meaning that $x, y \in S$ implies that $\frac{x+y}{2} \in S$. In practice, most sets we work with which are dyadic convex will also be convex, but in applications, it might be easier to check that S is merely dyadic convex.

[2] It is important that π only translates subsets by elements of 2Λ, so if $\pi\colon S \to 2P$ were injective, then we would have $\text{vol}(S) \leq \text{vol}(2P)$.

2.3.3 Sample Applications of Minkowski's Theorem

In order to convince us that Theorem 2.84 actually has a nontrivial use, we will provide a few sample applications before using it to show Theorem 2.102. All applications have essentially the same structure: embed a problem of interest into a lattice and then use Theorem 2.84 to find a small enough solution.

Our first application is from Diophantine approximation.

Proposition 2.87 (Dirichlet approximation)**.** Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ be an irrational number. Then for any positive integer $N > 1$, there exists a rational h/k with $0 < k \leq N$ such that

$$\left|\alpha - \frac{h}{k}\right| < \frac{1}{Nk}.$$

Proof using Theorem 2.84. We are looking for a pair of integers (h, k), so we might as well work with the lattice $\mathbb{Z}^2 \subseteq \mathbb{R}^2$ of rank 2. It remains to encode the needed bound to find a small solution. For technical reasons, fix some $\varepsilon > 0$ so that $N + \varepsilon < N + 1$.

Clearing fractions, we are being asked to look in the region

$$S \coloneqq \left\{(x, y) \in \mathbb{R} \times [-N - \varepsilon, N + \varepsilon] : |y\alpha - x| < \frac{1}{N}\right\}.$$

To apply Theorem 2.84, we have the following checks.

- Note that S is symmetric: if $(x, y) \in S$, then $(-x, -y) \in \mathbb{R} \times [-N - \varepsilon, N + \varepsilon]$ and $|(-y)\alpha - (-x)| = |y\alpha - x| < \frac{1}{N}$ verifies that $-(x, y) \in S$.
- We check that S is convex. If $(x_1, y_1) \in S$ and $(x_2, y_2) \in S$, then choose any $t \in [0, 1]$, and we use the triangle inequality to check that $y_1, y_2 \in [-N - \varepsilon, N + \varepsilon]$ implies $|ty_1 + (1 - t)y_2| \leq tN + (1 - t)N = N$ and
$$|(ty_1 + (1 - t)y_2)\alpha - (tx_1 + (1 - t)x_2)| \leq t\,|y_1\alpha - x_1| + (1 - t)\,|y_2\alpha - x_2| < \frac{1}{N}.$$
- We compute $\operatorname{vol} S$. Note that S is a parallelogram bounded by the line segments $y = -N - \varepsilon$ and $y = N + \varepsilon$ and $y\alpha - x = \frac{1}{N}$ and $y\alpha - x = -\frac{1}{N}$. Thus, the height (along the y-axis) is $2N + 2\varepsilon$, and the width of a parallel side is $2/N$, which multiplies to a volume of
$$(2N + 2\varepsilon) \cdot \frac{2}{N} > 4.$$

Combining the above checks, we see that Theorem 2.84 is able to provide a nonzero lattice point $(h, k) \in S \cap \mathbb{Z}^2$. In fact, note that $k \neq 0$: if $k = 0$, then $|-h| < 1/N$, forcing $h = 0$ too, which is not permitted. Now, by replacing (h, k) with $-(h, k)$ as necessary, we thus may assume that $k > 0$; additionally, $k \leq N + \varepsilon < N + 1$, so $k \leq N$ because k is an integer. Lastly, we see that

$$\left|\alpha - \frac{h}{k}\right| < \frac{1}{Nk}$$

by construction of (h, k). This completes the proof. ∎

Remark 2.88. Note that Proposition 2.87 follows from Proposition 1.40: let $\{h_n/k_n\}_{n=0}^{\infty}$ be the continued fraction convergents of α, and then we see $\{k_n\}_{n=0}^{\infty}$ is a strictly increasing sequence (e.g., by Proposition 1.32), so we may find n so that $k_n < N \leq k_{n+1}$. Then Proposition 1.40 implies

$$\left|\alpha - \frac{h_n}{k_n}\right| < \frac{1}{k_n k_{n+1}} \leq \frac{1}{Nk_n},$$

as desired.

Proposition 2.87 is a nice application, but we are here to solve quadratic equations, so let's give an application to solving quadratic equations.

Proposition 2.89. Fix an odd prime number p such that $p \equiv 1 \pmod 4$. Then there are integers (a, b) such that $p = a^2 + b^2$.

We will want the following number-theoretic lemma.

Lemma 2.90. Fix an odd prime number p such that $p \equiv 1 \pmod 4$. Then there is an integer x such that $x^2 \equiv -1 \pmod p$.

Proof. We proceed directly: we claim that $x \coloneqq \left(\frac{p-1}{2}\right)!$ will do the trick. Indeed, directly squaring, we see

$$\begin{aligned} x^2 &= 1 \cdot 2 \cdot 3 \cdot \ldots \cdot \frac{p-1}{2} \cdot \frac{p-1}{2} \cdot \ldots \cdot 3 \cdot 2 \cdot 1 \\ &\equiv (-1)^{(p-1)/2} \cdot 1 \cdot 2 \cdot 3 \cdot \ldots \cdot \frac{p-1}{2} \cdot -\frac{p-1}{2} \cdot \ldots \cdot -3 \cdot -2 \cdot -1 \\ &\equiv (p-1)!. \end{aligned}$$

To compute $(p-1)! \pmod p$, we pair off $x \in \{1, 2, \ldots, p-1\}$ off with the multiplicative inverse $x^{-1} \pmod p$; this pairing assigns x to a distinct unique element x^{-1} unless $x \equiv x^{-1}$, which is equivalent to $p \mid x^2 - 1$, or $x \equiv \pm 1 \pmod p$. Thus, everything outside ± 1 cancels out, so we are left with

$$x^2 \equiv (p-1)! \equiv 1 \cdot -1 \equiv -1 \pmod p,$$

as needed. ∎

Exercise 2.91. Fix an odd prime number p such that $p \equiv 3 \pmod 4$. Show that

$$\left(\left(\frac{p-1}{2}\right)!\right)^2 \equiv 1 \pmod p.$$

We are now ready to prove Proposition 2.89.

Proof of Proposition 2.89. The idea is to build a lattice Λ such that any $(a, b) \in \Lambda$ has $p \mid a^2 + b^2$ and then look for small vectors in Λ, hoping that we can show $a^2 + b^2 < 2p$. As such, we have the usual two steps.

1. We construct the desired lattice. It is here we will use that $p \equiv 1 \pmod 4$. By Lemma 2.90, there is $x \in \mathbb{Z}$ such that $x^2 \equiv -1 \pmod p$. We use this x to construct the lattice

$$\Lambda \coloneqq \left\{ a \begin{bmatrix} x \\ 1 \end{bmatrix} + b \begin{bmatrix} p \\ 0 \end{bmatrix} : a, b \in \mathbb{Z} \right\} = \left\{ a \begin{bmatrix} ax + bp \\ a \end{bmatrix} : a, b \in \mathbb{Z} \right\}.$$

   We now claim that any $(n, m) \in \Lambda$ has $p \mid n^2 + m^2$. Indeed, $(n, m) = (ax + bp, a)$ for some integers a and b, so

$$(ax + bp)^2 + a^2 \equiv a^2x^2 + a^2 \equiv a^2\left(x^2 + 1\right) \equiv 0 \pmod p.$$

   To apply Theorem 2.84, we will also want to compute $\operatorname{vol}\left(\mathbb{R}^2/\Lambda\right)$, which we see by Remark 2.75 is $\left|\det\left[\begin{smallmatrix} x & p \\ 1 & 0 \end{smallmatrix}\right]\right| = p$.

2. As discussed above, we want to set

$$S \coloneqq \left\{(x, y) : x^2 + y^2 < 2p\right\}.$$

   Here are our checks on S.

- Note that S is symmetric: $(x,y) \in S$ implies $x^2+y^2 < 2p$ and hence $(-x)^2+(-y)^2 < 2p$, so $-(x,y) \in S$.
- We check that S is convex. If $(x_1,y_1) \in S$ and $(x_2,y_2) \in S$, then choose any $t \in [0,1]$, and we use the triangle inequality to see

$$\|t(x_1,y_1)+(1-t)(x_2,y_2)\|_2 \le t\|(x_1,y_2)\|_2 + (1-t)\|(x_2,y_2)\|_2 = 2p,$$

  where $\|(x,y)\|_2 = \sqrt{x^2+y^2}$.
- Note that $\mathrm{vol}(S)$ is the area of our circle of radius $\sqrt{2p}$, which is simply $2p\pi$.

3. We now apply Theorem 2.84, which requires that we check $2p\pi > 4p$, which is true because $\pi > 2$. Thus, $\Lambda \cap S$ has a nonzero lattice point, which we label (a,b). Then the first step shows that $p \mid a^2+b^2$, but the construction of S requires $a^2+b^2 < 2p$. Because (a,b) is nonzero, we see $0 < a^2+b^2$ also, so we are left with $p = a^2+b^2$. ∎

Let's prove another result of this type.

Proposition 2.92. Fix a prime p such that there exists an integer x such that $x^2 \equiv -5 \pmod p$. Then there are integers (a,b) such that either $p = a^2+5b^2$ or $2p = a^2+5b^2$.

Proof. We imitate the proof of Proposition 2.89.

1. We construct the desired lattice. Fix our integer x with $x^2 \equiv -5 \pmod p$, and then we construct $\Lambda \subseteq \mathbb{R}^2$ as being spanned by $\left[\begin{smallmatrix} x \\ 1 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} p \\ 0 \end{smallmatrix}\right]$. Notably, any point on this lattice takes the form $(ax+bp, a)$ for some integers $a,b \in \mathbb{Z}$, so we see

$$(ax+bp)^2 + 5a^2 \equiv a^2\left(x^2+5\right) \equiv 0 \pmod p.$$

   As a last computation, we note that $\mathrm{vol}\left(\mathbb{R}^2/\Lambda\right) = \left|\det \left[\begin{smallmatrix} x & p \\ p & 0 \end{smallmatrix}\right]\right| = p$.

2. As suggested by the statement of the proposition, we set

$$S \coloneqq \left\{(x,y) : x^2+5y^2 < 3p\right\}.$$

   Here are our checks on S.

   - Note that S is symmetric: $(x,y) \in S$ implies $x^2+5y^2 < 3p$ and hence $(-x)^2+5(-y)^2 < 3p$, so $-(x,y) \in S$.
   - We check that S is convex. Pick up $(x_1,y_1),(x_2,y_2) \in S$ and $t \in [0,1]$, and we use the triangle inequality to see

$$\begin{aligned}\sqrt{(tx_1+(1-t)x_2)^2+5(ty_1+(1-t)y_2)^2} &= \left\|t(x_1,\sqrt{5}y_1)+(1-t)(x_2,\sqrt{5}y_2)\right\|_2 \\ &\le t\left\|(x_1,\sqrt{5}y_1)\right\|_2 + (1-t)\left\|(x_2,\sqrt{5}y_2)\right\|_2 \\ &< 3p.\end{aligned}$$

   - Note that $\mathrm{vol}(S)$ is the volume of an ellipse with one axis of length $\sqrt{3p}$ and the other axis of length $\sqrt{3p/5}$, so the area is

$$\frac{3}{\sqrt{5}} \cdot p \cdot \pi,$$

     which is greater than $4p$: indeed, it suffices for $3\pi > 4\sqrt{5}$, which is true because $(3\pi)^2 > 81 > 80$.

3. We now apply Theorem 2.84, which from the above checks provides nonzero $(a,b) \in \Lambda \cap S$, meaning that $p \mid a^2+5b^2$ and $0 < a^2+5b^2 < 3p$. This completes the proof. ∎

The above result is a bit unsatisfying because we have not determined which of $p = a^2 + 5b^2$ or $2p = a^2 + 5b^2$ is true. We will explain what is going on in more detail next week, specifically in Example 3.55.

It is now worth pointing out that the proof of Theorem 2.84 was inherently non-explicit: the proof essentially uses a pigeonhole principle, which promises us the nonzero lattice point without really telling us how to find it. As such, our above proof of Proposition 2.87 via Theorem 2.84 is actually less useful than the argument of Remark 2.88 because Remark 2.88 explains how to find the required rational promised in the statement.

It is a perfectly reasonable question to take $p \equiv 1 \pmod 4$ and then ask for the pair of integers (a, b) such that $p = a^2 + b^2$. This can in fact be efficiently computed (faster than brute-forcing values a where $1 \le a \le \sqrt{p/2}$), as we will discuss next week when we discuss quadratic forms.

2.3.4 Lattice Reduction

With the geometry of quadratic forms in place, we are able to contextualize the algorithm to make Theorem 2.84 constructive in the two-dimensional case for certain S. This is best seen by example.

Example 2.93. Let $p = 10037$ be prime. Given that $3271^2 \equiv -1 \pmod{10037}$, we find integers (x, y) such that $x^2 + y^2 = 10037$.

Proof. By the proof of Proposition 2.89, the shortest nonzero vector $\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]$ (with respect to $\|(x, y)\|^2 := x^2 + y^2$) in the lattice Λ spanned by $\left[\begin{smallmatrix} 3271 \\ 1 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 10037 \\ 0 \end{smallmatrix}\right]$ has magnitude less than $2p$ and will thus have $p = x^2 + y^2$. So we want to find short vectors in Λ. The idea is to imitate the Euclidean algorithm. We do this in steps.

1. We replace the longer vector $\left[\begin{smallmatrix} 10037 \\ 0 \end{smallmatrix}\right]$ with a shorter vector of the form $\left[\begin{smallmatrix} 10037 \\ 0 \end{smallmatrix}\right] - q\left[\begin{smallmatrix} 3271 \\ 1 \end{smallmatrix}\right]$. As long as q is an integer, the new pair of vectors will both in Λ and in fact span Λ. The Gram–Schmidt process would tell us to set $q = \left\langle \left[\begin{smallmatrix} 10037 \\ 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} 3271 \\ 1 \end{smallmatrix}\right] \right\rangle / \left\| \left[\begin{smallmatrix} 3271 \\ 1 \end{smallmatrix}\right] \right\|^2 = 3271/1066 \approx 3.1$ to project onto the line orthogonal to $\left[\begin{smallmatrix} 3271 \\ 1 \end{smallmatrix}\right]$, but as q must be an integer, we choose the closest integers as $q = 3$, yielding
$$\begin{bmatrix} 10037 \\ 0 \end{bmatrix} - 3\begin{bmatrix} 3271 \\ 1 \end{bmatrix} = \begin{bmatrix} 224 \\ -3 \end{bmatrix}.$$
2. We replace the longer vector $\left[\begin{smallmatrix} 3271 \\ 1 \end{smallmatrix}\right]$ with a shorter vector of the form $\left[\begin{smallmatrix} 3271 \\ 1 \end{smallmatrix}\right] - q\left[\begin{smallmatrix} 224 \\ -3 \end{smallmatrix}\right]$. Any integer q will suffice to continue to span Λ. Well, Gram–Schmidt suggests $q = \left\langle \left[\begin{smallmatrix} 3271 \\ 1 \end{smallmatrix}\right], \left[\begin{smallmatrix} 224 \\ -3 \end{smallmatrix}\right] \right\rangle / \left\| \left[\begin{smallmatrix} 224 \\ -3 \end{smallmatrix}\right] \right\|^2 = 73/5 = 14.6$, so we select the closest integer $q = 15$, yielding
$$\begin{bmatrix} 3271 \\ 1 \end{bmatrix} - 15\begin{bmatrix} 224 \\ -3 \end{bmatrix} = \begin{bmatrix} -89 \\ 46 \end{bmatrix}.$$

At this point, the pair $(-89, 46)$ has small enough coordinates that it is worth checking $89^2 + 46^2 = 10037$, so we are done. ∎

The above example might look needlessly complicated (the divisions involved are somewhat tedious), but it requires fewer total computations than the brute-force search, especially as p grows large. To work this out, try out Problem 3.1.7.

Let's generalize the process described above. It is worth comparing the result to Proposition 1.18.

Proposition 2.94 (Gaussian reduction)**.** Let $\langle \cdot, \cdot \rangle : \mathbb{R}^2 \to \mathbb{R}$ be a positive-definite inner product, and define $\|\cdot\| : \mathbb{R}^2 \to \mathbb{R}$ by $\|v\|^2 := \langle v, v \rangle$. Fix a lattice $\Lambda \subseteq \mathbb{R}$ of rank 2 spanned by v_0 with v_1 such that $\|v_0\| \ge \|v_1\|$, and define the sequence $v_2, v_3, \ldots$ of vectors and the sequence $q_2, q_3, \ldots$ of integers by
$$v_n = q_n v_{n+1} + v_{n+2},$$
where q_n is the closest integer to $p_n := \langle v_n, v_{n+1} \rangle / \|v_{n+1}\|^2$, breaking ties by using the integer closer to 0. Then there is some N such that $q_n = 0$ for $n \ge N$, and then a shortest vector in $\{v_N, v_{N+1}\}$ is a nonzero vector in Λ minimizing $\|\cdot\|$.

Proof. Quickly, note that, if Λ is spanned by the vectors v and w, then Λ is spanned by the vectors v and $w - cv$ for any integer c, so by induction, we see that Λ is spanned by the vectors v_n and v_{n+1} for any $n \geq 0$. In particular, $v_n \neq 0$ for all n.

We now proceed in steps.

1. We claim that $\|v_{n+2}\| \leq \|v_n\|$ for any n, with equality only when $q_n = 0$. This is a direct computation. We see that
$$\begin{aligned}\|v_{n+2}\|^2 &= \langle v_n - q_n v_{n+1}, v_n - q_n v_{n+1}\rangle \\ &= \|v_n\|^2 + q_n^2 \|v_{n+1}\|^2 - 2q_n\langle v_n, v_{n+1}\rangle,\end{aligned}$$
so it suffices to show that $q_n^2 \|v_{n+1}\|^2 \leq 2q_n\langle v_n, v_{n+1}\rangle$. If $q_n = 0$, there is nothing to say, so we take $|p_n| > 1/2$ in the following discussion.

   Now, if $\langle v_n, v_{n+1}\rangle > 0$, then $q_n > 0$, so it is enough to show that
$$q_n \overset{?}{<} 2 \cdot \frac{\langle v_n, v_{n+1}\rangle}{\|v_{n+1}\|^2}.$$
Well, q_n is a closest integer to p_n, and $p_n > 1/2$, so $q_n \leq p_n + 1/2 < 2p_n$, as desired. The argument in the case $\langle v_n, v_{n+1}\rangle < 0$ is essentially identical but merely adjusting signs.

2. We show that $q_n \neq 0$ only finitely often, establishing that there is some N for which $q_n = 0$ for $n \geq N$. Indeed, suppose there is an infinite set S in which $q_n \neq 0$ for each $n \in S$. Then $\{\|v_{n+2}\|\}_{n\in S}$ is an infinite strictly decreasing sequence by the previous step, so $\{v_{n+2}\}_{n\in S}$ is an infinite subset of Λ which should be bounded, which should contradict Proposition 2.77.

   To finish this step, we will show that there is some $R > 0$ such that $\{v \in \mathbb{R}^2 : \|v\| \leq \|v_0\|\}$ is contained in $[-R, R]^2$, which will complete the argument by Proposition 2.77 because $\|v_{n+2}\| \leq \|v_0\|$ for each $n \in S$. Let $\{e_1, e_2\}$ be an orthonormal basis for $\langle\cdot,\cdot\rangle$ constructed using (say) the Gram–Schmidt process, and let M be the matrix with e_1 and e_2 as columns. The point is that, letting $\|\cdot\|_2$ denote the Euclidean norm, we have
$$\|v\| = \left\|M^{-1}v\right\|_2.$$
It will suffice to find R such that $\left\|M^{-1}v\right\|_2 \leq \|v_0\|$ implies $\|v\|_2 \leq R$, or equivalently, $\|v\|_2 \leq \|v_0\|$ implies $\|Mv\|_2 \leq R$. Well, by continuity of M, there is some $\delta > 0$ such that $\|v\|_2 \leq \delta$ implies $\|Mv\|_2 \leq 1$, so $\|v\|_2 \leq \|v_0\|$ implies that $\left\|\frac{\delta}{\|v_0\|}v\right\|_2 \leq \delta$ and so
$$\|Mv\|_2 = \frac{\|v_0\|}{\delta} \cdot \left\|M\left(\frac{\delta}{\|v_0\|}v\right)\right\|_2 \leq \frac{\|v_0\|}{\delta},$$
so $R \coloneqq \|v_0\|/\delta$ will do.

3. We show that a shortest vector v_M in $\{v_N, v_{N+1}\}$ minimizes $\|\cdot\|$ among nonzero vectors in Λ. Namely, $M \in \{N, N+1\}$, and $v_N = v_{N+2}$ ensures $\|v_M\| \leq \|v_{M+1}\|$ in either case. The point is that $q_{M+1} = 0$ forces
$$\left|\frac{\langle v_M, v_{M+1}\rangle}{\|v_{M+1}\|^2}\right| \leq \frac{1}{2}.$$
Now, let $v \in \Lambda$ be a nonzero vector so that $v = av_M + bv_{M+1}$ for some integers $a, b \in \mathbb{Z}$ not both zero. Then we compute
$$\begin{aligned}\|v\|^2 &= \langle av_M + bv_{M+1}, av_N + bv_{M+1}\rangle \\ &= a^2\|v_M\|^2 + 2ab\langle v_M, v_{M+1}\rangle + b^2\|v_{M+1}\|^2 \\ &\geq a^2\|v_M\|^2 - ab\|v_M\|^2 + b^2\|v_M\|^2.\end{aligned}$$
Now, $[1,1,1]$ has discriminant -3, so it is a positive-definite quadratic form, so $a^2 - ab + b^2 > 0$ by Lemma 3.5. And because $a, b \in \mathbb{Z}$, we actually have $a^2 - ab + b^2 \geq 1$; we conclude $\|v\|^2 \geq \|v_M\|^2$, as desired. ∎

The benefit of working with a general inner product $\langle \cdot, \cdot \rangle$ is that we can now we can use Proposition 2.94 to reduce the value of our quadratic forms. As an example, we do a computation with Example 3.10.

Example 2.95. Let $p = 10039$ be prime. Given that $4691^2 \equiv -7 \pmod{p}$, we find integers (x, y) such that $x^2 + xy + 2y^2 = 10039$.

Proof. We track through the proof of example 3.10. Setting $x_0 := 4691$, we see that we want to set $x_1 := \frac{-1+4691}{2} = 2345$. Then the proof tells us that it suffices to find the "shortest" nonzero vector in the lattice Λ spanned by $\left[\begin{smallmatrix}2345\\1\end{smallmatrix}\right]$ and $\left[\begin{smallmatrix}10039\\0\end{smallmatrix}\right]$. Here, shortest is respect to the inner product $\langle \cdot, \cdot \rangle$ given by $[1, 1, 2]$; explicitly, $\langle (x_1, y_1), (x_2, y_2) \rangle := x_1x_2 + \frac{1}{2}(x_1y_2 + x_2y_1) + 2y_1y_2$. Anyway, we apply Proposition 2.94.

1. We compute $\left\langle \left[\begin{smallmatrix}10039\\0\end{smallmatrix}\right], \left[\begin{smallmatrix}2345\\1\end{smallmatrix}\right] \right\rangle / \left\| \left[\begin{smallmatrix}2345\\1\end{smallmatrix}\right] \right\|^2 = 4691/1096 \approx 4.3$, so we replace $\left[\begin{smallmatrix}10039\\0\end{smallmatrix}\right]$ by the vector
$$\begin{bmatrix}10039\\0\end{bmatrix} - 4\begin{bmatrix}2345\\1\end{bmatrix} = \begin{bmatrix}659\\-4\end{bmatrix}.$$
2. We compute $\left\langle \left[\begin{smallmatrix}2345\\1\end{smallmatrix}\right], \left[\begin{smallmatrix}659\\-4\end{smallmatrix}\right] \right\rangle / \left\| \left[\begin{smallmatrix}659\\-4\end{smallmatrix}\right] \right\|^2 = 307/86 \approx 3.6$, so we replace $\left[\begin{smallmatrix}2345\\1\end{smallmatrix}\right]$ by the vector
$$\begin{bmatrix}2345\\1\end{bmatrix} - 4\begin{bmatrix}659\\-4\end{bmatrix} = \begin{bmatrix}-291\\17\end{bmatrix}.$$
3. We compute $\left\langle \left[\begin{smallmatrix}659\\-4\end{smallmatrix}\right], \left[\begin{smallmatrix}-291\\17\end{smallmatrix}\right] \right\rangle / \left\| \left[\begin{smallmatrix}-291\\17\end{smallmatrix}\right] \right\|^2 = -37/16 \approx -2.3$, so we replace $\left[\begin{smallmatrix}659\\-4\end{smallmatrix}\right]$ by the vector
$$\begin{bmatrix}659\\-4\end{bmatrix} + 2\begin{bmatrix}-291\\17\end{bmatrix} = \begin{bmatrix}77\\30\end{bmatrix}.$$

From here, one could complete the lattice reduction (replacing $\left[\begin{smallmatrix}-291\\17\end{smallmatrix}\right]$ by $\left[\begin{smallmatrix}-137\\77\end{smallmatrix}\right]$ and then checking that we have finished our lattice reduction), but we are also able to see that 77 and 30 are reasonably small, so we can check $77^2 + 77 \cdot 30 + 2 \cdot 30^2 = 10039$, as desired. ∎

2.3.5 Problems

Do ten points worth of the following exercises.

Problem 2.3.1 (1 point). Find a basis $\{(x_1, y_1), (x_2, y_2)\}$ of the lattice $\mathbb{Z}^2 \subseteq \mathbb{R}^2$ such that $x_1, y_1, x_2, y_2 > 10$.

Problem 2.3.2 (2 points). Let $\Lambda \subseteq \mathbb{R}^n$ be a lattice. Show that there is a vector $v \in \Lambda \setminus \{0\}$ such that $\|v\|$ is minimized among all values in $\Lambda \setminus \{0\}$.

Problem 2.3.3 (3 points). Let $S^1 \subseteq \mathbb{C}^\times$ denote the subgroup of elements all of whose absolute values are 1, and let $G \subseteq S^1$ be a finite subgroup

(a) Consider the map $\pi \colon \mathbb{R} \to S^1$ given by $\pi(t) := \exp(2\pi i t)$. Show that $\pi^{-1}(G)$ is lattice in $\mathbb{R}$.

(b) Use (a) to show that G is cyclic.

(c) Use (b) to show that $\mu(\mathcal{O})$ is cyclic for any order $\mathcal{O}$ of a number field K.

Problem 2.3.4 (4 points). Fix a prime p such that there exists an integer x such that $x^2 \equiv -2 \pmod{p}$. Show that there is a pair of integers (a, b) such that $p = a^2 + 2b^2$.

Problem 2.3.5 (5 points)**.** The following problem requires the notion of a closed set: a subset $S \subseteq \mathbb{R}^n$ is said to be "closed" if and only if any convergent sequence $\{a_n\}_{n=0}^{\infty}$ contained in S has limit in S.

(a) (0 points) For experience, show that $[-1,1]^n \subseteq \mathbb{R}^n$ is a closed set for any positive integer n.

(b) (5 points) Let $\Lambda \subseteq \mathbb{R}^n$ be a lattice of rank n. For any closed, convex, symmetric about the origin subset $S \subseteq \mathbb{R}^n$ such that

$$\mathrm{vol}(S) \geq 2^n \,\mathrm{vol}\left(\mathbb{R}^n/\Lambda\right),$$

show that S contains a nonzero lattice point of Λ.

Problem 2.3.6 (5 points)**.** Let $\langle \cdot, \cdot \rangle \colon \mathbb{R}^2 \to \mathbb{R}$ be the standard inner product, and let $\Lambda \subseteq \mathbb{R}$ be a lattice of rank 2. Fixing $v_0, v_1 \in \Lambda$ as in the hypotheses of Proposition 2.94, define the sequence $v_2, v_3, \ldots$ of vectors and positive integer N as in Proposition 2.94.

(a) (2 points) Show that the (smaller) angle θ between the vectors v_N and v_{N+1} is at most $\pi/3$ (radians).

(b) (2 points) Show that the area of the parallelogram $\{av_N + bv_{N+1} : a, b \in [0,1]\}$ with sides v_N and v_{N+1} is $\|v_N\| \cdot \|v_{N+1}\| \cdot |\sin\theta|$. Conclude that

$$\mathrm{vol}\left(\mathbb{R}^2/\Lambda\right) \leq \frac{\sqrt{3}}{2} \|v_N\| \cdot \|v_{N+1}\|.$$

(c) (1 point) Find a lattice Λ where equalities are achieved in (a) and (b).

2.4 Dirichlet's Unit Theorem

In this section, we will prove the Dirichlet unit theorem, at long last.

2.4.1 Dirichlet's Unit Theorem: Set Up

We are almost at a point where we can state our main theorem. Approximately speaking, our goal is to generalize Proposition 2.7 to more general number fields. We now have enough machinery to explain where $x^2 - dy^2 = 1$ is coming from: given a non-square positive integer d, in the field $\mathbb{Q}(\sqrt{d})$, by Corollary A.35, the norm map $\mathrm{N}_{\mathbb{Q}(\sqrt{d})/\mathbb{Q}}$ is given by

$$\mathrm{N}_{\mathbb{Q}(\sqrt{d})/\mathbb{Q}}\left(x + y\sqrt{d}\right) = \left(x + y\sqrt{d}\right)\left(x - y\sqrt{d}\right) = x^2 - dy^2.$$

This also explains why we kept factoring $x^2 - dy^2$ into $\left(x - y\sqrt{d}\right)\left(x + y\sqrt{d}\right)$. It will shortly be helpful for us to have a more algebraic description of these elements.

Lemma 2.96. Fix a number field K. Then an element $u \in \mathcal{O}_K$ is a unit (i.e., has a multiplicative inverse in $\mathcal{O}_K$) if and only if $\left|\mathrm{N}_{K/\mathbb{Q}}(u)\right| = 1$.

Proof. We have two implications to show.

- Suppose that $u \in \mathcal{O}_K$ is a unit. Then we have some $v \in \mathcal{O}_K$ such that $uv = 1$. Taking norms of this equation, we see

  $$\mathrm{N}_{K/\mathbb{Q}}(u) \cdot \mathrm{N}_{K/\mathbb{Q}}(v) = 1.$$

  However, $\mathrm{N}_{K/\mathbb{Q}}(u), \mathrm{N}_{K/\mathbb{Q}}(v) \in \mathbb{Z}$ by Remark 2.39, and the only way to have two integers multiply to 1 is for them to be ± 1. Thus, $\mathrm{N}_{K/\mathbb{Q}}(u) = \pm 1$, as desired.

- Suppose that $\mathrm{N}_{K/\mathbb{Q}}(u) = \pm 1$ so that $\mathrm{N}_{K/\mathbb{Q}}(u)^2 = 1$. The point is to expand the norm using Corollary A.35 to get an equation of the form $uv = 1$. Indeed, by Corollary A.35, we see that

$$\prod_{i=1}^{n} \sigma_i(u)^2 = 1$$

where the $\sigma_1, \dots, \sigma_n \colon K \hookrightarrow \mathbb{C}$ are the embeddings of Proposition A.31. Identifying K with its image under $\sigma_1 \colon K \hookrightarrow \mathbb{C}$ (for example), we see that

$$\sigma_1(u) \cdot \underbrace{\sigma_1(u) \prod_{i=2}^{n} \sigma_i(u)^2}_{v :=} = 1.$$

Now, $uv = 1$, so we will be done once we establish that $v \in \mathcal{O}_K$. Well, $u \in \mathcal{O}_K$, so letting $f(x)$ be a monic polynomial with integer coefficients such that $f(u) = 0$, we see that $f(\sigma_i(u)) = 0$ for all i, so $\sigma_i(u)$ is an algebraic integer for each σ_i, so v is also an algebraic integer. Further, $v = 1/u \in K$, so it follows $v \in \mathcal{O}_K$. ■

So integer pairs (x, y) satisfying $x^2 - dy^2 = 1$ will be units in $\mathcal{O}_{\mathbb{Q}(\sqrt{d})}$. Note that the -1 case is also explained Lemma 2.96 because being a unit permits $\mathrm{N}_{K/\mathbb{Q}}(x + y\sqrt{d}) = -1$.

To continue, we observe that there is something a little off with Proposition 2.7. Namely, the proposition is only solving units in $\mathcal{O}^{\times}_{\mathbb{Q}(\sqrt{d})}$ of the form $x + y\sqrt{d}$ where $x, y \in \mathbb{Z}$, but we saw in Proposition 2.43 that sometimes we have a denominator of 2 present. Explicitly, one could use Proposition 2.7 to look for units in $\mathbb{Z}[\sqrt{5}]^{\times}$ even though $\mathbb{Z}[\sqrt{5}] \neq \mathcal{O}_{\mathbb{Q}(\sqrt{5})}$. As such, in the statement we prove, we will not want to only focus on the rings $\mathcal{O}_K$ but generalizes of them.

Definition 2.97 (order)**.** Fix a number field K of degree n over $\mathbb{Q}$. Then an *order* $\mathcal{O}$ is a subring of $\mathcal{O}_K$ which is a free abelian group of rank n. As with $\mathcal{O}_K$, we will abuse notation and write

$$\operatorname{disc} \mathcal{O} := \operatorname{disc}(\alpha_1, \dots, \alpha_n)$$

for any basis $\alpha_1, \dots, \alpha_n$ of $\mathcal{O}$. Note $\operatorname{disc} \mathcal{O}$ is still well-defined by Lemma 2.60.

Example 2.98. Fix a number field K. Then $\mathcal{O}_K$ is itself an order by Theorem 2.51.

Example 2.99. Let d be a non-square integer, and set $K := \mathbb{Q}(\sqrt{d})$. Then

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

is a subring of algebraic integers which is a free abelian group of rank 2 (with basis given by 1 and $\sqrt{d}$).

We are now almost ready to state our result. For technical reasons, we will want the notion of a signature.

Definition 2.100 (signature)**.** Fix a number field K of degree n over $\mathbb{Q}$, and let $\sigma_1, \dots, \sigma_n \colon K \hookrightarrow \mathbb{C}$ be the n embeddings of Proposition A.31.

- If an embedding $\sigma \colon K \hookrightarrow \mathbb{C}$ outputs to $\mathbb{R}$, we call σ a *real embedding*.
- Otherwise, $\sigma \colon K \hookrightarrow \mathbb{C}$ has output to $\mathbb{C} \setminus \mathbb{R}$ and is called a *complex embedding*.

Among the n embeddings $\sigma \colon K \hookrightarrow \mathbb{C}$, we let r_1 denote the number of real embeddings and $2r_2$ denote the number of complex embeddings, and we let (r_1, r_2) be the *signature* of K.

Remark 2.101. It is worth explaining why the number of complex embeddings $\sigma : K \hookrightarrow \mathbb{C}$ is even. Well, for any complex embedding $\sigma : K \hookrightarrow \mathbb{C}$, there is a complex conjugate $\overline{\sigma}(\alpha) \coloneqq \overline{\sigma(\alpha)}$ embedding. Because there is $\alpha \in K$ with $\sigma(\alpha) \in \mathbb{C} \setminus \mathbb{R}$, we see that $\overline{\sigma}(\alpha) \neq \sigma(\alpha)$, so $\overline{\sigma} \neq \sigma$, so these are in fact distinct embeddings. Thus, $\sigma \mapsto \overline{\sigma}$ defines a map from complex embeddings to complex embeddings, and $\overline{\overline{\sigma}} = \sigma$ implies that this is an involution, so it follows that the number of complex embeddings is even: we may pair a complex embedding σ off with its complex conjugate embedding!

At long last, here is our result.

Theorem 2.102 (Dirichlet unit)**.** Fix a number field K of signature (r_1, r_2). Let $\mu(K)$ denote the group of roots of unity in K. Let $\mathcal{O} \subseteq \mathcal{O}_K$ be an order, and let $\mu(\mathcal{O})$ be the roots of unity in $\mathcal{O}$. Then

$$\mathcal{O}^\times \cong \mu(\mathcal{O}) \times \mathbb{Z}^{r_1+r_2-1}.$$

In other words, there is a set of units $u_1, \dots, u_{r_1+r_2-1}$ such that, for any unit $u \in \mathcal{O}_K^\times$, there is a unique root of unity ζ and integers $n_1, \dots, n_{r_1+r_2-1}$ such that $u = \zeta u_1^{n_1} \cdots u_{r_1+r_2-1}^{n_{r_1+r_2-1}}$.

We are not going to prove Theorem 2.102 at all in this section; we will postpone until we have discussed a little Minkowski theory. For now, we satisfy ourselves with an example.

Example 2.103. Let's show that Theorem 2.102 appropriately generalizes Proposition 2.7. Fix a non-square positive integer d. Then $K \coloneqq \mathbb{Q}(\sqrt{d})$ has signature $(2, 0)$, and $\mu(K) = \{\pm 1\}$ because $\{\pm 1\}$ are the only roots of unity in $\mathbb{R}$. Thus, Theorem 2.102 implies that the order $\mathbb{Z}[\sqrt{d}]$ has

$$\mathbb{Z}[\sqrt{d}]^\times \cong \{\pm 1\} \times \mathbb{Z}.$$

Tracking Lemma 2.96 backward tells us that any solution $x^2 - dy^2 = \pm 1$ has $x + y\sqrt{d} = \pm \left(x_0 + y_0\sqrt{d}\right)^n$ for some unique sign $\pm$ and integer n. One can then reduce to $x^2 - dy^2 = 1$ as a subgroup of $\mathbb{Z}[\sqrt{d}]^\times$.

2.4.2 Dirichlet's Unit Theorem: Upper Bound

In this subsection, we prove what we can from Theorem 2.102 without using any Minkowski theory. The goal, roughly speaking, is to explain what the $r_1 + r_2 - 1$ is doing in the statement. In the discussion which follows, let K be a number field of degree n and signature (r_1, r_2), and we let $\rho_1, \dots, \rho_{r_1} : K \hookrightarrow \mathbb{C}$ denote the real embeddings, and we let $\sigma_1, \dots, \sigma_{r_2}$ be a subset of complex embeddings so that $\sigma_1, \dots, \sigma_{r_2}, \overline{\sigma_1}, \dots, \overline{\sigma_{r_2}}$ provides all complex embeddings. (See Remark 2.101.) Additionally, let $\mathcal{O}$ be an order.

The conclusion of Theorem 2.102 features the additive group $\mathbb{Z}$, but $\mathcal{O}^\times$ is a largely multiplicative object. We would thus like to turn our multiplicative problem and turn it into an additive one, which is done by taking $\log$s. To begin, the multiplicative problem we are interested in solving is essentially trying to ensure the equation

$$\prod_{i=1}^{r_1} |\rho_i(u)| \cdot \prod_{i=1}^{r_2} |\sigma_i(u)|^2 = 1,$$

which for $u \in \mathcal{O}_K$ we know is equivalent to $u \in \mathcal{O}_K^\times$ by Lemma 2.96 and Corollary A.35. To make this equation additive, we note that it is equivalent to

$$\sum_{i=1}^{r_1} \log |\rho_i(u)| + \sum_{i=1}^{r_2} \log |\sigma_i(u)|^2 = 0, \tag{2.4}$$

provided that $u \in K^\times$. Let's break down what just happened into two steps.

1. We use the embeddings to map K into some Euclidean space. With our enumeration, the most obvious thing to do is via the map $K \to \mathbb{R}^{r_1} \times \mathbb{C}^{r_2}$ given by $\alpha \mapsto (\rho_1(\alpha), \dots, \rho_{r_1}(\alpha), \sigma_1(\alpha), \dots, \sigma_{r_2}(\alpha))$.

However, we would like to work with real vector spaces, so we use the basis $\{1, i\}$ of $\mathbb{C}$ as an $\mathbb{R}$-vector space to define $j\colon K \to \mathbb{R}^n$ by

$$j\colon \alpha \mapsto (\rho_1(\alpha), \dots, \rho_{r_1}(\alpha), \operatorname{Re}\sigma_1(\alpha), \operatorname{Im}\sigma_1(\alpha), \dots, \operatorname{Re}\sigma_{r_2}(\alpha), \operatorname{Im}\sigma_{r_2}(\alpha)).$$

By construction, we see that $j\colon K \to \mathbb{R}^n$ is a homomorphism of additive groups.

2. After mapping $j\colon K \to \mathbb{R}^n$, we would like to take logarithms, so we define the map $\operatorname{Log}\colon (\mathbb{R}^\times)^n \to \mathbb{R}^{r_1+r_2}$ by

$$\operatorname{Log}(x_1, \dots, x_{r_1}, a_1, b_1, \dots, a_{r_2}, b_{r_2}) \coloneqq \left(\log|x_1|, \dots, \log|x_{r_2}|, \log\left|a_1^2 + b_1^2\right|, \dots, \log\left|a_{r_2}^2 + b_{r_2}^2\right|\right).$$

Observe that this is the same Log map that we saw in Proposition 2.19, and it will be useful for approximately the same reason: this map does a good job of measuring the "multiplicative" height of a nonzero element in $\mathcal{O}$.

Anyway, for $\alpha \in K^\times$, we have $\sigma(\alpha) \neq 0$ for each embedding σ, so $\operatorname{Log}(j(\alpha))$ is well-defined. And by construction (and by properties of $\log$), we see that the composite $(\operatorname{Log} \circ j)\colon K^\times \to \mathbb{R}^{r_1+r_2}$ is a homomorphism of groups.

If in addition $\alpha \in \mathcal{O}_K^\times$, then (2.4) tells us that $\operatorname{Log}(j(\alpha))$ lands in

$$H \coloneqq \left\{(x_1, \dots, x_{r_1+r_2}) : \sum_{i=1}^{r_1+r_2} x_i = 0\right\} \subseteq \mathbb{R}^{r_1+r_2},$$

which we call the "trace-0 hyperplane." Note that this is a hyperplane of $\mathbb{R}^{r_1+r_2}$ cut out by a single equation, so $\dim H = r_1 + r_2 - 1$. This is where the $r_1 + r_2 - 1$ in Theorem 2.102 will come from.

In order to justify our use of lattices, we note that $\mathcal{O} \subseteq K$ should in some sense feel like a "discrete subgroup" (compare this with $\mathbb{Z} \subseteq \mathbb{Q}$), so it is reasonable to expect that $j(\mathcal{O}) \subseteq \mathbb{R}^n$ and $\operatorname{Log}(j(\mathcal{O}^\times)) \subseteq H$ are discrete subgroups of Euclidean space and thus lattices. We show this now.

Proposition 2.104. Fix a number field K, and fix notation as above. Then $j(\mathcal{O}) \subseteq \mathbb{R}^n$ is a lattice of rank n with covolume $\operatorname{vol}(\mathbb{R}^n/j(\mathcal{O})) = \frac{1}{2^{r_2}}\sqrt{|\operatorname{disc}\mathcal{O}|}$.

Proof. By definition, $\mathcal{O}$ is a free abelian group of rank n, so produce a basis $\alpha_1, \dots, \alpha_n$. We claim that $j(\alpha_1), \dots, j(\alpha_n)$ provides a basis for $j(\mathcal{O}) \subseteq \mathbb{R}^n$. Certainly these elements span $j(\mathcal{O})$ because $j\colon \mathcal{O} \to \mathbb{R}^n$ is a homomorphism of additive groups, implying any $\alpha \in \mathcal{O}$ can be written as

$$j(\alpha) = j(c_1\alpha_1 + \cdots + c_n\alpha_n) = c_1 j(\alpha_1) + \cdots + c_n j(\alpha_n)$$

for some integers $c_1, \dots, c_n \in \mathbb{Z}$.

Now, to compute the covolume, we need to compute the determinant of the (transpose of the) matrix

$$\begin{bmatrix} \rho_1(\alpha_1) & \cdots & \rho_{r_1}(\alpha_1) & \operatorname{Re}\sigma_1(\alpha_1) & \operatorname{Im}\sigma_1(\alpha_1) & \cdots & \operatorname{Re}\sigma_{r_2}(\alpha_1) & \operatorname{Im}\sigma_{r_2}(\alpha_1) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_1(\alpha_n) & \cdots & \rho_{r_1}(\alpha_n) & \operatorname{Re}\sigma_1(\alpha_n) & \operatorname{Im}\sigma_1(\alpha_n) & \cdots & \operatorname{Re}\sigma_{r_2}(\alpha_n) & \operatorname{Im}\sigma_{r_2}(\alpha_n) \end{bmatrix}.$$

We would like to make this matrix look like the matrix for $\operatorname{disc}(\alpha_1, \dots, \alpha_n)$. Multiply each real part column by i times the imaginary column, which makes our determinant equal

$$\det \begin{bmatrix} \rho_1(\alpha_1) & \cdots & \rho_{r_1}(\alpha_1) & \sigma_1(\alpha_1) & \operatorname{Im}\sigma_1(\alpha_1) & \cdots & \sigma_{r_2}(\alpha_1) & \operatorname{Im}\sigma_{r_2}(\alpha_1) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_1(\alpha_n) & \cdots & \rho_{r_1}(\alpha_n) & \sigma_1(\alpha_n) & \operatorname{Im}\sigma_1(\alpha_n) & \cdots & \sigma_{r_2}(\alpha_n) & \operatorname{Im}\sigma_{r_2}(\alpha_n) \end{bmatrix}.$$

We now multiply each imaginary part column $\operatorname{Im}\sigma_i(\alpha_j)$ by $-2i$ and then add the corresponding $\sigma_i(\alpha_j)$ term to produce $\overline{\sigma_i}(\alpha_j)$, thus making our determinant

$$\frac{1}{(-2i)^{r_2}}\det\begin{bmatrix}\rho_1(\alpha_1) & \cdots & \rho_{r_1}(\alpha_1) & \sigma_1(\alpha_1) & \overline{\sigma_1}(\alpha_1) & \cdots & \sigma_{r_2}(\alpha_1) & \overline{\sigma_{r_2}}(\alpha_1)\\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ \rho_1(\alpha_n) & \cdots & \rho_{r_1}(\alpha_n) & \sigma_1(\alpha_n) & \overline{\sigma_1}(\alpha_n) & \cdots & \sigma_{r_2}(\alpha_n) & \overline{\sigma_{r_2}}(\alpha_n)\end{bmatrix}.$$

Taking absolute values, we see that this is $\frac{1}{2^{r_2}}\sqrt{|\operatorname{disc}(\alpha_1,\dots,\alpha_n)|}$, as needed. ■

Proposition 2.105. Fix a number field K, and fix notation as above. Then $\operatorname{Log}(j(\mathcal{O}^\times))\subseteq H$ is a lattice.

Proof. The point is to apply Proposition 2.77. Fix any $M>0$, and we show that $\operatorname{Log}(j(\mathcal{O}^\times))\cap[-M,M]^{r_1+r_2}$ is finite, which will be enough by Proposition 2.77 because we already know that $\operatorname{Log}(j(\mathcal{O}^\times))\subseteq H$ from the discussion above.

Well, we simply pull back to $\mathbb{R}^n$. Indeed, $\operatorname{Log}(x_1,\dots,x_{r_1},a_1,b_1,\dots,a_{r_2},b_{r_2})\in[-M,M]^n$ implies that $|x_\bullet|\le e^M$ and $|a_\bullet|,|b_\bullet|<e^{M/2}$ always. Thus, $\operatorname{Log}(j(\alpha))\in[-M,M]^n$ implies that

$$j(\alpha)\in[-\exp(M),\exp(M)]^n,$$

but only finitely many α satisfy this by Proposition 2.77 because $j(\mathcal{O})\subseteq\mathbb{R}^n$ is a lattice by Proposition 2.104. ■

We know that $\operatorname{Log}(j(\mathcal{O}^\times))$ is a lattice, so it is a free abelian group of rank at most $\dim H = r_1+r_2-1$. However, the statement of Theorem 2.102 includes some roots of unity. Where did they go?

Lemma 2.106. Fix a number field K, and fix notation as above. Then $\ker(\operatorname{Log}\circ j)$ is the set $\mu(\mathcal{O})$ of roots of unity in $\mathcal{O}$, and $\mu(\mathcal{O})$ is finite.

Proof. Quickly, note that some $\alpha\in\mathcal{O}_K^\times$ lives in $\ker(\operatorname{Log}\circ j)$ if and only if $|\sigma(\alpha)|=1$ for all embeddings $\sigma\colon K\hookrightarrow\mathbb{C}$. Let $S\subseteq\mathcal{O}^\times$ be the set of such α. In one direction, certainly $\mu(\mathcal{O})\subseteq S$: for any $\zeta\in\mu(\mathcal{O})$, we have $\zeta^n=1$ for some n, so $|\sigma(\zeta)|^n=1$ and thus $|\sigma(\zeta)|=1$ for all embeddings $\sigma\colon K\hookrightarrow\mathbb{C}$.

For the other direction, we will show that S is finite; this will complete the proof because $S\subseteq\mathcal{O}^\times$ is a subgroup, meaning that any $\alpha\in S$ has $\alpha^{\#S}=1$ and thus $\alpha\in\mu(\mathcal{O})$. Anyway, to show that S is finite, we use the proof of Proposition 2.105: note that any α in the kernel must have $|\sigma(\alpha)|=1$ for all embeddings $\sigma\colon K\hookrightarrow\mathbb{C}$. But then

$$\ker(\operatorname{Log}\circ j)\subseteq j(\mathcal{O})\cap[-1,1]^n$$

is finite by Proposition 2.104, so we are done. ■

Remark 2.107. One can more directly show that any $\alpha\in\mathcal{O}_K$ such that $|\sigma(\alpha)|=1$ for all embeddings $\sigma\colon K\hookrightarrow\mathbb{C}$ must be a root of unity. Here is the argument. Let $S\subseteq\mathcal{O}_K$ be the set of such α. For each α, let $f(x)\in\mathbb{Z}[x]$ be the corresponding monic irreducible polynomial (see Lemma 2.38) so that

$$f(x)=\prod_{\sigma\colon\mathbb{Q}(\alpha)\hookrightarrow\mathbb{C}}(x-\sigma(\alpha)).$$

However, a direct expansion reveals that the x^r coefficient of this polynomial is the sum of $\binom{d}{d-r}$ complex numbers of absolute value 1, where $d\le n$ is the degree of $f(x)$. Thus, the set of polynomials $f(x)\in\mathbb{Z}[x]$ which can correspond to some $\alpha\in S$ is finite, so S is finite. However, we can see that $S\subseteq K^\times$ is a subgroup, so it follows that $\alpha^{\#S}=1$ for any $\alpha\in S$.

We now combine Lemma 2.106 with Proposition 2.105 to achieve the result, which is quite close to Theorem 2.102.

Proposition 2.108. Fix a number field K, and fix notation as above. Let r be the rank of the lattice $\mathrm{Log}(j(\mathcal{O}^\times)) \subseteq H$. Then

$$\mathcal{O}^\times \cong \mu(\mathcal{O}^\times) \times \mathbb{Z}^r.$$

Proof. Let $v_1, \dots, v_r \in \mathrm{Log}(j(\mathcal{O}^\times))$ be a basis, and find any $\alpha_1, \dots, \alpha_r \in \mathcal{O}^\times$ so that $\mathrm{Log}(j(\alpha_i)) = v_i$ for each i. Then we define the function

$$\varphi \colon \mu(\mathcal{O}^\times) \times \mathbb{Z}^r \to \mathcal{O}^\times$$

by $\varphi(\zeta, (e_1, \dots, e_r)) \coloneqq \zeta \alpha_1^{e_1} \cdots \alpha_r^{e_r}$. We claim that φ is a group isomorphism, which will complete the proof. Here are our checks.

- Homomorphism: we check

$$\varphi\left(\zeta\zeta', (e_1 + e_1', \dots, e_r + e_r')\right) = \zeta\zeta' \cdot \alpha_1^{e_1+e_1'} \cdots \alpha_r^{e_r+e_r'} = \varphi\left(\zeta, (e_1, \dots, e_r)\right) \varphi\left(\zeta', (e_1', \dots, e_r')\right).$$

- Injective: suppose $\varphi(\zeta, (e_1, \dots, e_r)) = 1$. Then

$$\alpha_1^{e_1} \cdots \alpha_r^{e_r} = \zeta^{-1},$$

  so passing through $\mathrm{Log} \circ j$ tells us that $e_1 v_1 + \cdots + e_r v_r = 0$ by Lemma 2.106, so $(e_1, \dots, e_r) = (0, \dots, 0)$ because the $v_\bullet$ are linearly independent. But then the above equation implies $\zeta = 1$ as well.

- Surjective: fix $u \in \mathcal{O}^\times$. Then $\mathrm{Log}(j(u)) \in \mathrm{Log}(j(\mathcal{O}^\times))$ has some $(e_1, \dots, e_r) \in \mathbb{Z}^r$ such that

$$\mathrm{Log}(j(u)) = e_1 v_1 + \cdots + e_r v_r.$$

  But then $\mathrm{Log}\left(j(u\alpha_1^{-e_1} \cdots \alpha_r^{-e_r})\right) = 0$, so $u\alpha_1^{-e_1} \cdots \alpha_r^{-e_r} \in \mu(\mathcal{O})$, so $u = \zeta \alpha_1^{e_1} \cdots \alpha_r^{e_r}$ for some $\zeta \in \mu(\mathcal{O})$. ∎

2.4.3 Dirichlet's Unit Theorem: Lower Bound

We continue with the notation of the previous subsection; for example, K is a number field, and $\mathcal{O} \subseteq \mathcal{O}_K$ is an order. From Proposition 2.108, it remains to compute the rank of the lattice $\mathrm{Log}(j(\mathcal{O}^\times)) \subseteq H$.

Proposition 2.109. Fix a number field K and an order $\mathcal{O} \subseteq \mathcal{O}_K$, and fix notation as in section 2.4.2. Then $\mathrm{Log}(j(\mathcal{O}^\times)) \subseteq H$ is a lattice of rank $r_1 + r_2 - 1$.

Note Theorem 2.102 would then follow immediately from Propositions 2.108 and 2.109. It remains to prove Proposition 2.109, which is the goal of this subsection. This requires exhibiting many units. For this, our argument will be similar in spirit to Proposition 2.13: we will produce lots of elements of small norm, and we will use quotients of these in order to produce the required units.

To begin, we need to know that there are not many elements of small norm.

Lemma 2.110. Fix a number field K and an order $\mathcal{O} \subseteq \mathcal{O}_K$. For any nonzero $\alpha \in \mathcal{O}$, we have $[\mathcal{O} : (\alpha)] = \left|\mathrm{N}_{K/\mathbb{Q}}(\alpha)\right|$.

Proof. By definition, $\mathrm{N}_{K/\mathbb{Q}}(\alpha)$ is the determinant of the multiplication-by-α map $\mu_\alpha \colon \mathcal{O} \to \mathcal{O}$, whose image is exactly (α). But Lemma 2.76 then tells us that $|\det \mu_\alpha| = [\mathcal{O} : (\alpha)]$, so the result follows. ∎

Lemma 2.111. Fix a number field K and an order $\mathcal{O} \subseteq \mathcal{O}_K$. For any positive integer N, there are only finitely many ideals $I \subseteq \mathcal{O}$ such that $[\mathcal{O} : I] = N$.

Proof. In fact, there are only finitely many additive subgroups I of $\mathcal{O} \cong \mathbb{Z}^n$ of index N. Well, any subgroup $I \subseteq \mathbb{Z}^n$ of index N makes the quotient $\mathbb{Z}^n/I$ annihilated by N, so $NI \supseteq \mathbb{Z}^n$. Thus, $N\mathbb{Z}^n \subseteq I \subseteq \mathbb{Z}^n$, so I may be recovered by its image $I/N\mathbb{Z}^n \subseteq \mathbb{Z}^n/N\mathbb{Z}^n$, but $\mathbb{Z}^n/N\mathbb{Z}^n$ is a finite group, so there are only finitely many options for I. ■

Proposition 2.112. Fix a number field K and an order $\mathcal{O} \subseteq \mathcal{O}_K$. For any positive integer N, there are only finitely many $\alpha \in \mathcal{O}/\mathcal{O}^\times$ such that $\left|\mathrm{N}_{K/\mathbb{Q}}(\alpha)\right| = N$.

Proof. Quickly, note $\left|\mathrm{N}_{K/\mathbb{Q}}(\alpha)\right|$ is well-defined for $\alpha \in \mathcal{O}/\mathcal{O}^\times$ because $\alpha \in \mathcal{O}_K^\times$ is equivalent to $\left|\mathrm{N}_{K/\mathbb{Q}}(\alpha)\right| = 1$ by Lemma 2.96.

Now, we note that $\alpha \cdot \mathcal{O}^\times = \alpha' \cdot \mathcal{O}^\times$ if and only if $(\alpha) = (\alpha')$ by tracking through our principal ideals. Thus, Lemma 2.110 tells us that we are asking for finitely (principal) ideals $I \subseteq \mathcal{O}$ such that $[\mathcal{O} : I] = N$. This finiteness follows from Lemma 2.111. ■

We now use Proposition 2.112 to produce units. To begin, we need many elements of small norm.

Lemma 2.113. Fix a number field K and an order $\mathcal{O} \subseteq \mathcal{O}_K$, and fix notation as in section 2.4.2. Further, fix an index $1 \le i_0 \le r_1 + r_2$. Then there is an absolute constant $C(\mathcal{O}) > 0$ with the following property: for any nonzero $\alpha \in \mathcal{O}$, there is a nonzero $\beta \in \mathcal{O}$ such that $\left|\mathrm{N}_{K/\mathbb{Q}}(\alpha')\right| \le C(\mathcal{O})$ and writing

$$\mathrm{Log}(j(\alpha)) = (a_1, \dots, a_{r_1+r_2}) \qquad \text{and} \qquad \mathrm{Log}(j(\alpha')) = (a_1', \dots, a_{r_1+r_2}')$$

requires $a_i' < a_i$ for each $i \neq i_0$.

Proof. We use Theorem 2.84. Fix $C(\mathcal{O}) > 0$ to be determined later. Our lattice will be $j(\mathcal{O}) \subseteq \mathbb{R}^n$, which we know to be of rank n by Proposition 2.104. To achieve $a_i' < a_i$, we define $S \subseteq \mathbb{R}^n$ by

$$S \coloneqq \left\{(x_1, \dots, x_n) : |x_1| < e^{c_1}, \dots, |x_{r_1}| < e^{c_{r_1}}, \left|x_{r_1+1}^2 + b_{r_1+2}^2\right| < e^{c_{r_1+1}}, \dots, \left|x_{n-1}^2 + x_n^2\right| < e^{c_{r_1+r_2}}\right\},$$

where we require $c_i = a_i$ for each $i \neq i_0$ and

$$e^{c_{i_0}} \coloneqq \frac{C(\mathcal{O})}{\prod_{i \neq i_0} e^{c_i}}.$$

Notably, S is the product of r_1 intervals and r_2 circles, so its volume is

$$\mathrm{vol}(S) = 2^{r_1} e^{c_1 + \cdots + c_{r_1}} \cdot \pi^{r_2} e^{c_{r_1+1} \cdots c_{r_1+r_2}} = 2^{r_1} \pi^{r_2} C(\mathcal{O}).$$

For $C(\mathcal{O})$ big enough, we see $2^{r_1}\pi^{r_2}C(\mathcal{O}) > \mathrm{vol}(\mathbb{R}^n/j(\mathcal{O}))$ (note that $C(\mathcal{O})$ only needs to depend on $\mathcal{O}$), so Theorem 2.84 yields a nonzero $j(\alpha') \in j(\mathcal{O})$ such that $\alpha' \in S$. Looking at the construction of S, we see that the inequalities on $\mathrm{Log}(j(\alpha'))$ hold by construction (look at $c_i = a_i$ for $i \neq i_0$), and $\left|\mathrm{N}_{K/\mathbb{Q}}(\alpha')\right| \le C(\mathcal{O})$ again hold by construction (look at c_{i_0}). ■

And here are our units.

Lemma 2.114. Fix a number field K and an order $\mathcal{O} \subseteq \mathcal{O}_K$, and fix notation as in section 2.4.2. For any index $1 \le i_0 \le r_1 + r_2$, there is a unit $\gamma \in \mathcal{O}^\times$ such that writing

$$\mathrm{Log}(j(\gamma)) = (u_1, \dots, u_{r_1+r_2})$$

has $u_i < 0$ for $i \neq i_0$. In fact, $u_{i_0} > 0$.

Proof. Quickly, note that having $u_i < 0$ for $i \neq i_0$ implies $u_{i_0} > 0$ because the coordinates need to sum to 0 because $\operatorname{Log}(j(u)) \subseteq H$.

Anyway, fix $C(\mathcal{O})$ as in Lemma 2.113. We now inductively use Lemma 2.113 to produce many elements of norm bounded by $C(\mathcal{O})$, which will be required to give us units by Proposition 2.112. To begin, we may assume $C(\mathcal{O}) > 1$, and we set $\alpha_1 := 1$. Then we apply Lemma 2.113 on α_1 to produce some α_2 with norm at most $C(\mathcal{O})$ as well, and then we do this again to produce α_3, and so on.

This produces an infinite list of elements of norm at most $C(\mathcal{O})$, but the set of classes in $\mathcal{O}/\mathcal{O}^\times$ represented by such elements is finite by Proposition 2.112, so we may find distinct elements of the sequence α and α' which differ by a unit so that $\alpha\gamma = \alpha'$. We claim that u is the desired unit: by construction of the sequence, we see that writing

$$\begin{aligned}\operatorname{Log}(j(\alpha)) &= (a_1, \dots, a_{r_1+r_2}) \\ \operatorname{Log}(j(\alpha')) &= (a_1', \dots, a_{r_1+r_2}') \\ \operatorname{Log}(j(\gamma)) &= (u_1, \dots, u_{r_1+r_2})\end{aligned}$$

requires $a_i' < a_i$ for each $i \neq i_0$, so we are done upon noting $u_i = a_i' - a_i$ for each i. ∎

Applying Lemma 2.114 to each unit, we produce units $\gamma_1, \dots, \gamma_{r_1+r_2} \in \mathcal{O}^\times$ such that exactly the ith component of the $\operatorname{Log}(j(\gamma_i))$ is positive. Now, to prove Proposition 2.109, it is enough to produce $r_1 + r_2 - 1$ linearly independent vectors, meaning we want to show

$$\operatorname{rank} \begin{bmatrix} | & & | \\ \operatorname{Log}(j(\gamma_1)) & \cdots & \operatorname{Log}(j(\gamma_{r_1+r_2-1})) \\ | & & | \end{bmatrix} = r_1 + r_2 - 1.$$

It turns out to be easier to consider the transpose matrix, whereupon the result becomes the following piece of linear algebra.

Lemma 2.115. Let $A = (a_{ij})_{1 \leq i,j, \leq n}$ be a matrix such that the rows sum to zero and $a_{ij} < 0$ for each $i \neq j$ and $a_{ii} > 0$ for each i. Then $\operatorname{rank} A = n - 1$.

Proof. Certainly $\operatorname{rank} A < n$ because $(1, \dots, 1) \in \mathbb{R}^{n \times n}$ is in the kernel. To establish $\operatorname{rank} A \geq n - 1$, we will show that the first $n - 1$ columns of A are linearly independent. Enumerate the columns of A as $v_1, \dots, v_n$. Any nontrivial linear relation among the first $n - 1$ of these columns

$$\sum_{i=1}^{n-1} c_i v_i = 0$$

may find some $|c_{i_0}| > 0$ as large as possible and then divide the entire equation by c_{i_0} so that the new equation has $c_{i_0} = 1$ while $|c_i| \leq 1$ for each i. However, row $i_0 < n$ now has

$$0 = \sum_{i=1}^{n-1} c_i a_{i_0 i} \geq \sum_{i=1}^{n} a_{i_0 i} > \sum_{i=1}^{n} a_{i_0 i} = 0,$$

which is a contradiction. ∎

Proposition 2.109 now follows from the lemma and the discussion immediately preceding it. This completes the proof of Theorem 2.102.

2.4.4 A Harder Problem Revisited

We are now equipped to understand and prove Proposition 2.22. The following lemma will be useful.

Lemma 2.116. Fix a cubic number field K with exactly one real embedding $\rho\colon K \hookrightarrow \mathbb{R}$. For any order $\mathcal{O} \subseteq \mathcal{O}_K$, we have $\mathcal{O}^\times \cong \mu(\mathcal{O}) \times \mathbb{Z}$, and any unit $u \in \mathcal{O}^\times$ with $\rho(u) > 1$ will in fact have

$$\rho(u)^3 + 7 > \frac{1}{4}\left|\operatorname{disc}\mathcal{O}\right|.$$

Proof. The first claim is a direct consequence of Theorem 2.102: having exactly one real embedding implies that the signature (r_1, r_2) is $(1, r_2)$ where $1 + 2r_2 = [K : \mathbb{Q}] = 3$, meaning the signature is actually $(r_1, r_2) = (1, 1)$. Then $r_1 + r_2 - 1 = 1$, so we are done by Theorem 2.102.

It remains to show the inequality. The point is that $u \in \mathcal{O}$ implies that $\left|\operatorname{disc}(1, u, u^2)\right| \geq \left|\operatorname{disc}\mathcal{O}\right|$ by Lemma 2.50. To compute $\operatorname{disc}(1, u, u^2)$, let $\sigma\colon K \hookrightarrow \mathbb{C}$ be a complex embedding so that our embeddings are $\rho, \sigma, \overline{\sigma}\colon K \hookrightarrow \mathbb{C}$. Define $r \coloneqq 1/\sqrt{\rho(u)} < 1$ so that $|\sigma(u)|^2 = \left|\mathrm{N}_{K/\mathbb{Q}}(u)\right|/\rho(u) = r^2$ (we are using Lemma 2.96) and so $\sigma(u) = re^{i\theta}$ for some θ. As such, we compute

$$\begin{aligned}
\operatorname{disc}\left(1, u, u^2\right) &= \det\begin{bmatrix} 1 & \rho(u) & \rho(u)^2 \\ 1 & \sigma(u) & \sigma(u)^2 \\ 1 & \overline{\sigma}(u) & \overline{\sigma}(u)^2 \end{bmatrix}^2 \\
&= \left(\left(\sigma(u)\overline{\sigma}(u)^2 - \sigma(u)^2\overline{\sigma}(u)\right) - \left(\rho(u)\overline{\sigma}(u)^2 - \rho(u)^2\sigma(u)\right) + \left(\rho(u)\sigma(u)^2 - \rho(u)^2\sigma(u)\right)\right)^2 \\
&= \left((\rho(u) - \sigma(u))(\sigma(u) - \overline{\sigma}(u))(\overline{\sigma}(u) - \rho(u))\right)^2 \\
&= \left(\sigma(u) - \overline{\sigma}(u)\right)^2 \left|\rho(u) - \sigma(u)\right|^2 \\
&= -4r^2(\sin\theta)^2 \left|1/r^2 - r\cos\theta - ir\sin\theta\right|^4 \\
&= -4r^2(\sin\theta)^2 \left((1/r^2 - r\cos\theta)^2 + (r\sin\theta)^2\right)^2 \\
&= -4r^2(\sin\theta)^2 \left(1/r^4 - 2(1/r)\cos\theta + r^2\right)^2 \\
&= -4(\sin\theta)^2 \left(r^3 + 1/r^3 - 2\cos\theta\right)^2.
\end{aligned}$$

This is negative, so we take absolute values to see

$$\frac{1}{4}\left|\operatorname{disc}\mathcal{O}\right| < (\sin\theta)^2\left(r^3 + \frac{1}{r^3} - 2\cos\theta\right)^2.$$

It remains to bound the right-hand side. To manipulate, set $c \coloneqq \cos\theta$ and $s \coloneqq r^3 + 1/r^3$; note $s \geq 2$ because $r^3 + 1/r^3 - 2 \geq \left(r^3 - 1/r^3\right) \geq 0$. As such,

$$\begin{aligned}
(\sin\theta)^2\left(r^3 + \frac{1}{r^3} - 2\cos\theta\right) &= \left(1 - c^2\right)(s - 2c)^2 \\
&= \left(1 - c^2\right)\left(s^2 - 4sc + 4c^2\right) \\
&= s^2 - 4sc + 4c^2 - s^2c^2 + 4sc^3 - 4c^4 \\
&= s^2 + 4 - \left(sc + 2 - 2c^2\right)^2.
\end{aligned}$$

Thus,

$$\frac{1}{4}\left|\operatorname{disc}\mathcal{O}\right| < \left(r^3 + \frac{1}{r^3}\right)^2 + 4 = r^6 + \frac{1}{r^6} + 6 < \rho(u)^3 + 7,$$

as needed. ∎

Proposition 2.117. The cubic number field $K \coloneqq \mathbb{Q}(\sqrt[3]{2})$ has exactly one real embedding, and $\mathcal{O} \coloneqq \mathbb{Z}[\sqrt[3]{2}]$ is an order. The element $u \coloneqq 1 + \sqrt[3]{2} + \sqrt[3]{4}$ is a unit, and any element of $\mathcal{O}^\times$ can be written uniquely in the form $\pm u^n$ for some integer n.

Proof. By the argument of Proposition A.31, $K \cong \mathbb{Q}[x]/\left(x^3-2\right)$ has the real embedding $\sqrt[3]{2} \mapsto \sqrt[3]{2}$ and two complex embeddings $\sqrt[3]{2} \mapsto e^{2\pi i/3}\sqrt[3]{2}$ and $\sqrt[3]{2} \mapsto e^{4\pi i/3}\sqrt[3]{2}$. In the argument which follows, we identify K with its embedding in $\mathbb{R}$. To finish up the claims of the first sentence, we see that $\mathcal{O}$ has basis given by $\left\{1, \sqrt{3}[2], \sqrt[3]{4}\right\}$: these certainly generate $\mathcal{O}$, and they are a $\mathbb{Q}$-linearly independent basis of K, so they are certainly linearly independent over $\mathbb{Z}$.

Next, we note that u is in fact a unit because $\mathrm{N}_{K/\mathbb{Q}}(u) = 1+2+4-6 = 1$ by Example 2.40. It remains to show that any other unit in $\mathcal{O}$ can be written in the form $\pm u^n$. This follows by Theorem 2.102 and Lemma 2.116. Because $\mathcal{O} \subseteq \mathbb{R}$, we see that $\mu(\mathcal{O}) = \{\pm 1\}$. (All other roots of unity in $\mathbb{C}$ do not live in $\mathbb{R}$.) It remains to find a unit $u_0 \in \mathcal{O}^\times$ such that

$$\mathcal{O}^\times = \{\pm 1\} \times u_0^{\mathbb{Z}},$$

which exists by Theorem 2.102. By adjusting the sign of u_0, we may assume $u_0 > 0$. By replacing u_0 with $1/u_0$ as needed, we may assume $u_0 > 1$.

Now, we see that we have $u = u_0^n$ for some positive integer n. We claim that $u = u_0$, which will complete the proof. It is enough to show that $u < u_0^2$, or equivalently, $u^3 < u_0^6$. Well, recall $\operatorname{disc}\mathcal{O} = -108$, so $u_0^3 > \frac{1}{4}\cdot 108 - 7 = 20$. But now $u^3 < (1+2+4)^3 = 343 < 400$, so $u^3 < u_0^6$, so we are done. ∎

At long last, Proposition 2.117 follows from Proposition 2.22.

Proposition 2.22. Define the sequence of ordered triples of nonnegative integers $\{(x_0, y_0, z_0)\}_{n\in\mathbb{Z}}$ recursively by $(x_0, y_0, z_0) \coloneqq (1, 0, 0)$ and

$$(x_{n+1}, y_{n+1}, z_{n+1}) = (x_n + 2y_n + 2z_n, x_n + y_n + 2z_n, x_n + y_n + z_n)$$

for any $n \in \mathbb{Z}$. Then for any triple (x, y, z) of integers, we have $x^3 + 2y^3 + 4z^3 - 6xyz = 1$ if and only if $(x, y, z) = (x_n, y_n, z_n)$ for integer n.

Proof. By Example 2.40, any solution to $x^3 + 2y^3 + 4z^3 - 6xyz = 1$ is really a norm-1 unit of $x + y\sqrt[3]{2} + z\sqrt[3]{4}$. Looking at the units provided by Proposition 2.117, we see that the norm-1 units are of the form

$$x_n + y_n\sqrt[3]{2} + z_n\sqrt[3]{4} = (1 + \sqrt[3]{2} + \sqrt[3]{4})^n$$

for an integer $n \in \mathbb{Z}$. Explicitly, the sign -1 has norm -1, and $\mathrm{N}_{\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q}}(1 + \sqrt[3]{2} + \sqrt[3]{4}) = 1$, so the units of the form $(1 + \sqrt[3]{2} + \sqrt[3]{4})^n$ have norm 1, and the units of the form $-(1 + \sqrt[3]{2} + \sqrt[3]{4})^n$ have norm -1. To finish, and the recursion provided by Proposition 2.117 exactly describes the needed triples (x_n, y_n, z_n), so we are done! ∎

2.4.5 Problems

Do ten points worth of the following exercises.

Problem 2.4.1 (3 points). Let $S^1 \subseteq \mathbb{C}^\times$ denote the subgroup of elements all of whose absolute values are 1, and let $G \subseteq S^1$ be a finite subgroup

(a) Consider the map $\pi\colon \mathbb{R} \to S^1$ given by $\pi(t) \coloneqq \exp(2\pi i t)$. Show that $\pi^{-1}(G)$ is lattice in $\mathbb{R}$.

(b) Use (a) to show that G is cyclic.

(c) Use (b) to show that $\mu(\mathcal{O})$ is cyclic for any order $\mathcal{O}$ of a number field K.

Problem 2.4.2 (3 points). Let Λ' and Λ be lattices in $\mathbb{R}^n$ of rank n. Suppose $\Lambda' \subseteq \Lambda$.

(a) Show that there is a positive integer m such that $m \operatorname{vol}(\mathbb{R}^n/\Lambda) = \operatorname{vol}(\mathbb{R}^n/\Lambda')$.

(b) If $\operatorname{vol}(\mathbb{R}^n/\Lambda) = \operatorname{vol}(\mathbb{R}^n/\Lambda')$, show that $\Lambda = \Lambda'$.

Problem 2.4.3 (4 points)**.** Classify the integer solutions to

$$x^3 + 3y^3 + 9z^3 - 9xyz = 1$$

in a way akin to Proposition 2.22.

THEME 3

QUADRATIC EQUATIONS: FACTORIZATION

3.1 Binary Quadratic Forms

At the start of this unit, we were quick to discard equations of the form

$$ax^2 + bxy + cy^2 = n$$

where $b^2 - 4ac < 0$ because these can be checked via a finite computation. However, we will see that there is still interesting structure to uncover here.

3.1.1 Geometry of Quadratic Forms

Here is our definition of interest.

Definition 3.1 (binary quadratic form)**.** Given three integers numbers $a, b, c \in \mathbb{Z}$, we may define the *binary quadratic form* as

$$f(x, y) \coloneqq ax^2 + bxy + cy^2.$$

We may abbreviate this to $[a, b, c]$.

Example 3.2. We have the binary quadratic form $[1, 0, 1]$, which is $x^2 + y^2$.

Quadratic forms are typically defined via real symmetric matrices: given a binary quadratic form $f = [a, b, c]$, we see that

$$f(x, y) = ax^2 + bxy + cy^2 = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} a & b/2 \\ b/2 & c \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix},$$

so it will be convenient to define $M_f = M_{[a,b,c]} \coloneqq \left[\begin{smallmatrix} a & b/2 \\ b/2 & c \end{smallmatrix}\right]$. The determinant of this matrix is of interest.

Definition 3.3 (discriminant)**.** Fix a quadratic form $f = [a, b, c]$. Then we define the *discriminant* of f to be $\operatorname{disc} f \coloneqq -4 \det M_f = b^2 - 4ac$. If $\operatorname{disc} f < 0$ and $a > 0$, we say that f is *positive definite*.

Example 3.4. We have $\operatorname{disc}[1, 0, 1] = -4$.

In reality, the adjective "positive definite" is one given to inner products, so let's explain this. Given a positive-definite quadratic form $f = [a, b, c]$, we see that M_f is a real symmetric matrix of positive determinant, so we define

$$\langle (x_1, y_1), (x_2, y_2) \rangle_f \coloneqq \begin{bmatrix} x_1 & y_1 \end{bmatrix} \begin{bmatrix} a & b/2 \\ b/2 & c \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = ax_1x_2 + \frac{b}{2}(x_1y_2 + y_1x_2) + cy_1y_2. \tag{3.1}$$

In particular, the actual quadratic form f is the squared norm of $\langle \cdot, \cdot \rangle_f$. Let's check that this provides a positive-definite inner product.

Lemma 3.5. Fix a positive-definite binary quadratic form $f = [a, b, c]$. Then $\langle \cdot, \cdot \rangle_f$ is a positive-definite inner product on $\mathbb{R}^2$. In particular,

$$f(x, y) = a\left(x + \frac{b}{2a}y\right)^2 - \frac{b^2 - 4ac}{4a}y^2.$$

Proof. We run our checks one at a time. Fix $(x_1, y_1), (x_2, y_2), (x_3, y_3) \in \mathbb{R}^2$ and some $r \in \mathbb{R}$.

- Symmetric: we see that $\langle (x_1, y_1), (x_2, y_2) \rangle_f = \langle (x_2, y_2), (x_1, y_1) \rangle_f$ directly from (3.1).
- Linear: we see that

$$\begin{aligned} \langle r(x_1, y_1) + (x_2, y_2), (x_3, y_3) \rangle_f &= a(rx_1 + x_2)x_3 + \frac{b}{2}((rx_1 + x_2)y_3 + (ry_1 + y_2)x_3) + c(ry_1 + y_2)y_3 \\ &= r\left(ax_1x_3 + \frac{b}{2}(x_1y_3 + y_1x_3) + cy_1y_3\right) \\ &\quad + \left(x_2x_3 + \frac{b}{2}(x_2y_3 + y_2x_3) + cy_2y_3\right) \\ &= r\langle (x_1, y_1), (x_3, y_3) \rangle_f + \langle (x_2, y_2), (x_3, y_3) \rangle_f. \end{aligned}$$

- Positive-definite: fix nonzero $(x, y) \in \mathbb{R}^2$. We took $a > 0$, so we complete the square to write

$$ax^2 + bxy + cy^2 = a\left(x^2 + \frac{b}{a}xy + \frac{b^2}{4a^2}y^2\right) - \frac{b^2 - 4ac}{4a}y^2 = a\left(x + \frac{b}{2a}y\right)^2 - \frac{b^2 - 4ac}{4a}y^2.$$

  The right-hand side is a linear combination of squares with positive coefficients, so it will be positive as long as one of the squares is nonzero. Well, if both squares vanish, then $y = 0$ and $x + \frac{b}{2a}y = 0$ and so $x = 0$ too, but $(x, y) \neq (0, 0)$ by assumption. ∎

We now appeal to facts about inner products to discuss the geometry given by f.

Lemma 3.6 (Cauchy–Schwarz)**.** Fix a positive-definite inner product $\langle \cdot, \cdot \rangle \colon V \to \mathbb{R}$ on a real vector space V. For any $v, w \in \mathbb{R}^2$, we see

$$|\langle v, w \rangle|^2 \leq \langle v, v \rangle \langle w, w \rangle.$$

Proof. If $v = 0$, then both sides vanish. Otherwise, we make take $v \neq 0$ so that $\|v\| \neq 0$. For brevity, write $\|u\|^2 \coloneqq \langle u, u \rangle$.

1. We claim that

$$\left\| \|v\|^2 w - \langle v, w \rangle v \right\|^2 = \|v\|^4 \|w\|^2 - \|v\|^2 \langle v, w \rangle^2.$$

   This is a direct expansion. For brevity, write $a \coloneqq \|v\|^2$ and $b \coloneqq \langle v, w \rangle$. Then

$$\begin{aligned} \|aw - bv\|^2 &= \langle aw - bv, aw - bv \rangle \\ &= a^2 \|w\|^2 - 2ab\langle v, w \rangle + b^2 \|v\|^2 \\ &= a^2 \|w\|^2 - 2ab^2 + ab^2, \end{aligned}$$

   which is what we wanted upon simplifying and plugging in for a and b.

2. To complete the proof, we see that
$$\|v\|^2 \|w\|^2 - \langle v, w\rangle^2 \geq \frac{1}{\|v\|^2} \left\| \|v\|^2 w - \langle v, w\rangle v \right\|^2 \geq 0,$$
so we are done. ∎

Proposition 3.7 (triangle inequality)**.** Fix a positive-definite inner product $\langle \cdot, \cdot \rangle : V \to \mathbb{R}$ on a real vector space V. For any $v, w \in \mathbb{R}^2$, we see
$$\sqrt{\langle v + w, v + w\rangle} \leq \sqrt{\langle v, v\rangle} + \sqrt{\langle w, w\rangle}.$$

Proof. Note that these square roots are defined by Lemma 3.5, which tells us that $f(v) = \langle v, v\rangle_f \geq 0$ for any $v \in \mathbb{R}^2$.

Anyway, we may directly expand
$$\langle v + w, v + w\rangle = \langle v, v\rangle + \langle w, w\rangle + 2\langle v, w\rangle.$$
By Lemma 3.6, $\langle v, w\rangle^2 \leq \langle v, v\rangle\langle w, w\rangle$, so the right-hand side is bounded above by $\left(\sqrt{\langle v, v\rangle} + \sqrt{\langle w, w\rangle}\right)^2$, so we are done upon taking square roots everywhere. ∎

Corollary 3.8. Fix a positive-definite inner product $\langle \cdot, \cdot \rangle : V \to \mathbb{R}$ on a real vector space V. For any $C > 0$, the set
$$\left\{v \in \mathbb{R}^2 : \langle v, v\rangle < C\right\}$$
is convex and symmetric about the origin.

Proof. Let the given set be S. Anyway, we do our checks separately.

- Convex: we use Proposition 3.7. Indeed, fix any $v, w \in S$ and $t \in [0, 1]$. Then
$$\sqrt{\langle tv + (1-t)w, tv + (1-t)w\rangle} \leq \sqrt{\langle tv, tv\rangle} + \sqrt{\langle (1-t)w, (1-t)w\rangle} < tC + (1-t)C = C.$$
- Symmetric about the origin: if $v \in S$, then $\langle -v, -v\rangle = \langle v, v\rangle < C$, so $-v \in C$. ∎

Let's see some applications. We begin by generalizing Proposition 2.92.

Proposition 3.9. Fix a positive-define quadratic form $f = [a, b, c]$ of discriminant d. Let n be an odd integer coprime to a such that there exists an integer x_0 such that $x_0^2 \equiv d \pmod{n}$. Then there are integers (x, y) such that $mn = f(x, y)$ for some positive integer $n \leq 2\sqrt{-d}/\pi$.

Proof. We follow the proof of Proposition 2.92.

1. We construct the desired lattice. Quickly, we claim that there is x_1 such that $ax_1^2 + bx_1 + c \equiv 0 \pmod{n}$. By the quadratic formula, we want
$$x_1 \equiv \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \equiv \frac{-b \pm x_0}{2a} \pmod{n},$$
which is solvable. Note that these fractions with denominator $2a$ are legal because $\gcd(n, 2a) = 1$.

   Using our given x_1, we choose $\Lambda \subseteq \mathbb{R}^2$ as being spanned by $\left[\begin{smallmatrix} x_1 \\ 1 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} n \\ 0 \end{smallmatrix}\right]$. Any point on this lattice takes the form $(a_0 x_1 + b_0 n, a_0)$, so we see
$$f(a_0 x_1 + b_0 n, a_0) \equiv a a_0^2 x_1^2 + b a_0^2 x + c \equiv 0 \pmod{n}.$$

2. We construct the desired set as
$$S \coloneqq \{(x,y) : f(x,y) < Nn\},$$
for N to be set later. Note that S is convex and symmetric about the origin by Corollary 3.8. It remains to compute the volume. Expanding out with Lemma 3.5, we see
$$S = \left\{(x,y) : \left(\sqrt{a}x + \frac{b}{2\sqrt{a}}y\right)^2 + \left(\frac{\sqrt{-d}}{2\sqrt{a}}y\right)^2 < Nn\right\},$$
which becomes the circle of radius $\sqrt{Nn}$ upon applying
$$\begin{bmatrix} \sqrt{a} & 0 \\ b/(2\sqrt{a}) & \sqrt{-d}/(2\sqrt{a}) \end{bmatrix}.$$
Thus, the area is $Nn\pi$ upon applying the above linear transformation of determinant $d/2$, so the volume of S is $2Nn\pi/\sqrt{-d}$.

3. We now apply Theorem 2.84, which requires us to check that $2Nn\pi/\sqrt{-d} > 4n$, which is equivalent to $N > 2\sqrt{-d}/\pi$, so we set $N \coloneqq 2\sqrt{-d}/\pi + \varepsilon$ for small $\varepsilon > 0$. Thus, we get a nonzero pair $(x_\varepsilon, y_\varepsilon) \in \Lambda \cap S$, so $f(x_\varepsilon, y_\varepsilon)$ is divisible by p and $0 < f(x_\varepsilon, y_\varepsilon) < Nn + \varepsilon$, where the left inequality holds because f is positive-definite (by Lemma 3.5).

   We now claim that we have some nonzero $(x,y) \in \Lambda \cap S$ with $0 < f(x,y) < Nn$, which will complete the proof. Indeed, suppose not. Then sending $\varepsilon \to 0^+$ in the previous paragraph produces an infinite sequence $\{(x_\varepsilon, y_\varepsilon)\}_{\varepsilon = 1/n, n \in \mathbb{Z}^+}$ of points with $f(x_\varepsilon, y_\varepsilon)$ descending to ε. But Λ can only have finite intersection with any given S, so we have a contradiction. ■

Example 3.10. Consider $f = [1,1,2]$, which has discriminant -7, and we see $2 > 2\sqrt{7}/\pi$ because $\pi^2 > 9 > 7$. Thus, using Proposition 3.9 with $n = 2$, for any odd prime p such that there exists an integer x_0 such that $x_0^2 \equiv -7 \pmod{p}$, we get an integer pair (x,y) such that $p = f(x,y) = x^2 + xy + 2y^2$.

Example 3.11. Consider $f = [1,1,5]$, which has discriminant -19, and we see $3 > 2\sqrt{19}/\pi$ because $\pi^2 > 9 > 19 \cdot 4/9$. Thus, using Proposition 3.9 with $n = 3$, for any odd prime p such that there exists an integer x_0 such that $x_0^2 \equiv -19 \pmod{p}$, we get an integer pair (x,y) such that $p = f(x,y) = x^2 + xy + 5y^2$.

3.1.2 Equivalence of Forms

In many cases, Section 2.3.4 and the proof of Proposition 3.9 tells how to solve an equation of the form $ax^2 + bxy + cy^2 = n$ provided that there is a solution. It still remains to determine when a solution exists. In order to be able to provide a more robust method than what Minkowski's geometry of numbers provides, we need to study binary quadratic forms on their own terms.

Approximately speaking, we are most interested in the numbers represented by a binary quadratic form $[a,b,c]$, and essentially only this information. We should perhaps give these equations a name.

Definition 3.12 (represents)**.** A binary quadratic form f *represents* an integer n if and only if there are integers $(x,y) \in \mathbb{Z}^2$ such that $f(x,y) = n$. If we require in addition that $\gcd(x,y) = 1$, then we say that f *properly represents* n.

Roughly speaking, we care about properly representing an integer n because they describe what is "new" to f representing an integer. Namely, if $\gcd(x,y) > 1$, then we could just factor out $d \coloneqq \gcd(x,y)$ to see that $f(x,y) = n$ is really arising from $f(x/d, y/d) = n/d^2$.

Example 3.13. The binary quadratic form $f(x,y) = x^2 + y^2$ represents 25 because $0^2 + 5^2 = 25$ but also properly represents 25 because $3^2 + 4^2 = 25$.

Example 3.14. The binary quadratic form $f(x,y) = x^2 + y^2$ represents 49 because $0^2 + 7^2 = 49$, but it does not properly represent 49. Indeed, if $x^2 + y^2 = 49$ with $\gcd(x,y) = 1$, then in particular 7 cannot divide either x nor y, so we see that $x^2 + y^2 \equiv 0 \pmod 7$ implies

$$\left(\frac{x}{y}\right)^2 \equiv -1 \pmod 7,$$

which is impossible because then x/y would be an element of order 4 in the group $(\mathbb{Z}/7\mathbb{Z})^\times$, which has order 6.

Example 3.15. Suppose that a binary quadratic form $f = [a,b,c]$ represents a prime number p. Then f actually properly represents p: indeed, if $ax^2 + bxy + cy^2 = p$, then any common factor d of x and y will have $d^2 \mid p$. For example, $d \mid p$, so $d \in \{\pm 1, \pm p\}$, but $p^2 \nmid p$, so we must have $d \in \{\pm 1\}$, meaning that $\gcd(x,y) = 1$.

In the sequel, we will mostly focus on properly representing integers because representing primes will always be proper by Example 3.15.

Akin to not wanting to care about $\gcd(x,y) > 1$ in solutions to $f(x,y) = n$ where $f = [a,b,c]$, we will also want to not care about $\gcd(a,b,c) > 1$ for the same reason. So here is an adjective to fix that.

Definition 3.16 (primitive)**.** A binary quadratic form $f = [a,b,c]$ is *primitive* if and only if $\gcd(a,b,c) = 1$.

Remark 3.17. Note that $f = [a,b,c]$ has $f(1,0) = a$ and $f(1,1) = a + b + c$ and $f(0,1) = c$. Thus, f is primitive if and only if $\gcd(f(1,0), f(1,1), f(0,1)) = 1$. One can make our definition of primitive even more intrinsic to the function f by not caring about the basis; see Problem 3.1.3.

Anyway, because we are only interested in determining if a binary quadratic form represents some integer, we may as well allow ourselves some freedom in swapping our binary quadratic forms.

Definition 3.18 (equivalent)**.** Two binary quadratic forms $f = [a,b,c]$ and $g = [a',b',c']$ are *equivalent*, written $f \sim g$, if and only if there is $M \in \mathrm{SL}_2(\mathbb{Z})$ such that

$$M_f = M^\intercal M_g M.$$

Remark 3.19. An immediate benefit to our matrix view of binary quadratic forms is that we can quickly see that $f \sim g$ implies that $\operatorname{disc} f = \operatorname{disc} g$: with $M \in \mathrm{SL}_2(\mathbb{Z})$ such that $M_f = M^\intercal M_g M$, we see

$$\operatorname{disc} f = -4 \det M_f = -4(\det M^\intercal \cdot \det M_g \cdot \det M) = -4 \det M_g = \operatorname{disc} g.$$

Remark 3.20. It is a little more tedious to check that any form equivalent to a primitive one remains primitive, but it is true. Suppose that $f = [a, b, c]$ is equivalent to $g = [a', b', c']$, and we show that $\gcd(a, b, c)$ divides $\gcd(a', b', c')$; this is enough because equivalence is symmetric (see Lemma 3.21). Well, we are promised $M := \left[\begin{smallmatrix} p & q \\ r & s \end{smallmatrix}\right]$ so that $M_f = M^\intercal M_g M$. Thus, for any $x, y \in \mathbb{Z}$, we see that

$$f(x, y) = \begin{bmatrix} x & y \end{bmatrix} M^\intercal M_g M \begin{bmatrix} x \\ y \end{bmatrix} = g(px + qy, rx + sy)$$

is divisible by $\gcd(a', b', c')$. Thus, $\gcd(a, b, c) = \gcd(f(1, 0), f(1, 1), f(0, 1))$ (see Remark 3.17) is divisible by $\gcd(a', b', c')$.

We should probably check that this is an equivalence relation.

Lemma 3.21. Equivalence is an equivalence relation on the set of binary quadratic forms.

Proof. Here are our checks. Fix binary quadratic forms f, g, h.

- Reflexive: note that $M_f = I_2^\intercal M_f I_2$, so $f \sim f$ follows.
- Symmetric: if $f \sim g$, then we have $M \in \mathrm{SL}_2(\mathbb{Z})$ such that $M_f = M^\intercal M_g M$, so $M_g = \left(M^{-1}\right)^\intercal M_f M^{-1}$, so $g \sim f$.
- Transitive: if $f \sim g$ and $g \sim h$, then we get $M, N \in \mathrm{SL}_2(\mathbb{Z})$ such that $M_f = M^\intercal M_g M$ and $M_g = N^\intercal M_h N$, so
$$M_f = M^\intercal M_g M = M^\intercal N^\intercal M_h N M = (NM)^\intercal M_h (NM),$$
so $f \sim h$ follows. ∎

Let's check that equivalence does what we want it to do.

Lemma 3.22. Fix two binary quadratic forms f and g such that there is $M \in \mathbb{Z}^{2\times 2}$ with $|\det M| = 1$ with

$$M_f = M^\intercal M_g M.$$

Then, for any integer n, the f (properly) represents n if and only if g (properly) represents n.

Proof. The hypothesis on M tells us that M is invertible and $M^{-1} \in \mathbb{Z}^{2\times 2}$. Now, in one direction, if f represents n if and only if there is a vector $v := \left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right] \in \mathbb{Z}^2$ such that $v^\intercal M_f v = n$, which implies

$$v^\intercal M_f v = v^\intercal M^\intercal M_g M v = (Mv)^\intercal M_g (Mv),$$

so Mv witnesses g representing n. Furthermore, if f properly represents n, then we may assume $\gcd(x, y) = 1$, then we would like to show that Mv also has coprime coordinates. Well, set $M := \left[\begin{smallmatrix} p & q \\ r & s \end{smallmatrix}\right]$ with $pq - rs = \det M = \pm 1$ so that

$$Mv = \begin{bmatrix} px + qy \\ rx + sy \end{bmatrix}$$

where $ps - qr = \pm 1$. But then $\gcd(px + qy, rx + sy)$ divides $\gcd(prx + qry, prx + psy) = \gcd(prx + qry, psy - qry) = \gcd(prx + qry, y)$ and so divides y; furthermore, $\gcd(px + qy, rx + sy)$ will divide $\gcd(pdx + qsy, qrx + qsy) = \gcd(psx - qrx, qrx + qsy) = \gcd(x, qrx + qsy)$ and so divides x also. So $\gcd(px + qy, rx + sy) = 1$.

Finishing up, we note that the reverse direction is similar, merely replacing M with $M^{-1} \in \mathbb{Z}^{2\times 2}$ because $M_g = \left(M^{-1}\right)^\intercal M_f M^{-1}$. ∎

Remark 3.23. Lemma 3.22 suspiciously allows for M of determinant -1 when checking for our binary quadratic forms to (properly) represent the same integers. As such, one might reasonably want to adjust our definition of equivalence to allow for M with $\det M = -1$. It turns out that this makes the theory a bit harder to handle for reasons not immediately apparent.

Let's reap some quick benefit from our only caring about binary quadratic forms up to equivalence.

Proposition 3.24. Fix a binary quadratic form f. Then f properly represents an integer $n \in \mathbb{Z}$ if and only if $f \sim [n, b', c']$ for some integers $b', c' \in \mathbb{Z}$.

Proof. If $f \sim [n, b', c']$, then we are done by Lemma 3.22: $[n, b', c']$ properly represents n because $n \cdot 1^2 + b' \cdot 1 \cdot 0 + c' \cdot 0^2 = n$.

In the reverse direction, suppose that $f = [a, b, c]$ is equivalent to $[n, b', c']$. Then there is a matrix $M := \left[\begin{smallmatrix} p & q \\ r & s \end{smallmatrix}\right]$ such that

$$\begin{bmatrix} p & r \\ q & s \end{bmatrix} \begin{bmatrix} a & b/2 \\ b/2 & c \end{bmatrix} \begin{bmatrix} p & q \\ r & s \end{bmatrix} = \begin{bmatrix} n & b'/2 \\ b'/2 & c \end{bmatrix}.$$

Computing just the top-left corner of the left-hand side reveals that $f(p, r) = n$, and further $\gcd(p, r) = 1$ because $pq - rs = 1$. ∎

Remark 3.25. Note that Proposition 3.24 is non-constructive: even if we are told b' and c' so that $f \sim [n, b', c']$, it is not obvious how to go back and construct the matrix M witnessing this equivalence. Nonetheless, we ought to be able to go and back solve $f(x, y) = n$ using the methods of section 3.1.4.

Corollary 3.26. Let D be an integer which is either 0 or $1 \pmod 4$. Then an odd integer n is properly represented by a primitive binary quadratic form f of discriminant D if and only if there is an integer b such that $b^2 \equiv D \pmod n$.

Proof. If n is properly represented by a primitive binary quadratic form f of discriminant D, then Proposition 3.24 combined with Lemma 3.22 allows us to assume that $f = [n, b, c]$ for some integers $b, c \in \mathbb{Z}$. But then $D = b^2 - 4nc$, so $b^2 \equiv D \pmod n$.

Conversely, suppose that $b^2 \equiv D \pmod n$ for some integer b; by replacing b with $n - b$ as needed, we may assume that b and n have the same parity because n is odd. Then we may write $D = b^2 - nc_0$ for some integer c_0, but taking $\pmod 4$ reveals that $D \equiv b^2 \pmod 4$, so in fact $4 \mid c_0$, so $D = b^2 - 4nc$ for some integer c. But then $[n, b, c]$ is a binary quadratic form of discriminant D which of course properly represents n (by, say, Proposition 3.24). Note that $[n, b, c]$ is primitive because $\gcd(n, b, c)$ divides $\gcd\left(n, b^2 - 4nc\right)$ but $\gcd(n, D) = 1$. ∎

Example 3.27. Let p be an odd prime, and let D be an integer which is either 0 or $1 \pmod 4$. By Example 3.15, p is properly represented by a binary quadratic form of discriminant D if and only if p is merely represented by a binary quadratic form of discriminant D, which by Corollary 3.26 is equivalent to having some $b \in \mathbb{Z}$ with $b^2 \equiv D \pmod p$.

Corollary 3.26 now motivates us to show that there are relatively few binary quadratic forms of given discriminant, up to equivalence. In fact, we will shortly show that there are finitely many, but it is possible to have more than one. To see this, we want to be able to put binary quadratic forms into a reasonably canonical "reduced" form.

3.1.3 Reduced Forms

As promised at the end of the previous subsection, we take the following notion of reduced.

Definition 3.28 (reduced)**.** A binary quadratic form $f = [a, b, c]$ with negative discriminant is *semi-reduced* if and only if $|b| \le a \le c$. If in addition we have $b \ge 0$ if $|b| = a$ or $a = c$, then we say that f is *reduced*.

Example 3.29. The binary quadratic form $[1, 0, n]$, or $x^2 + ny^2$, is reduced for any positive integer n. Similarly, for any positive integer $n \equiv 3 \pmod 4$, the binary quadratic form

$$x^2 + xy + \frac{n+1}{4}y^2$$

is reduced.

Here is the relevant theorem.

Theorem 3.30. Every binary quadratic form of negative discriminant is equivalent to a unique reduced form.

Proof. This proof has two parts: we show that any binary quadratic form $f = [a, b, c]$ with $\operatorname{disc} f < 0$ is equivalent to some reduced form, and then we show that no two distinct reduced forms are equivalent. For the first part, we note that we have the two moves

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a & b/2 \\ b/2 & c \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} c & -b/2 \\ -b/2 & a \end{bmatrix} \tag{3.2}$$

$$\begin{bmatrix} 1 & 0 \\ m & 1 \end{bmatrix} \begin{bmatrix} a & b/2 \\ b/2 & c \end{bmatrix} \begin{bmatrix} 1 & m \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} a & ma + b/2 \\ ma + b/2 & c(a, b, c, m) \end{bmatrix}, \tag{3.3}$$

where $c(a, b, c, m) = m(ma + b) + c$ is some integer. These moves will allow us to transform f into a reduced form.

Quickly, we note that we may adjust by $\left[\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right]$ to enforce $a \ge 0$, but because $b^2 - 4ac < 0$, we now need to have $c \ge 0$ as well. Now, we may apply (3.2) to update $f = [a, b, c]$ to have $a \le c$. After, we may apply (3.3) to update $f = [a, b, c]$ via the Euclidean algorithm to enforce $-a/2 \le b/2 \le a/2$, or $|b| \le a$. It is possible that applying (3.2) will make it so that $a > c$ (because c changed), but then we can apply (3.2) again to get back to $a \le c$. Of course, we might now no longer have $|b| \le a$, but then we reapply (3.3), and repeat the process. Note that this will terminate eventually because the value of $a \ge 0$ is strictly decreasing on each application of (3.2).

In total, we have gotten f to be equivalent to a semi-reduced form. To get f to be reduced, we have to deal with the sign of b.

- If $a = c$, then (3.2) does not adjust a or c, but it will replace b with $-b$, allowing us to assume that $b \ge 0$.
- If $b = -a$, then (3.3) with $m = 1$ does not adjust a or c at all, but it will replace $b = -a$ with $b = a$.

The above points have transformed f into a reduced form, completing the first part of the argument.

For the second part of the argument, suppose that $f = [a, b, c]$ and $f' = [a', b', c']$ are equivalent primitive binary quadratic forms of the same discriminant. Our goal is to show $a = a'$ and $b = b'$ and $c = c'$; without loss of generality, take $a \ge a'$. Now, we are given integers $p, q, r, s \in \mathbb{Z}$ such that

$$\begin{bmatrix} p & r \\ q & s \end{bmatrix} \begin{bmatrix} a & b/2 \\ b/2 & c \end{bmatrix} \begin{bmatrix} p & q \\ r & s \end{bmatrix} = \begin{bmatrix} a' & b'/2 \\ b'/2 & d' \end{bmatrix}.$$

For example, $a' = ap^2 + bpr + cr^2$. Now, the main point is the bound

$$a \ge a' = ap^2 + bpr + cr^2 \ge ap^2 - a\,|pr| + ar^2 \ge a\,|pr|\,, \tag{3.4}$$

where the rightmost inequality holds because it is equivalent to $a\,(|p| - |r|)^2 \ge 0$. Thus, $|pr| \le 1$, so $p, r \in \{-1, 0, 1\}$. We have the following cases.

- Note $(p,r)=(0,0)$ would imply that $\operatorname{disc} f'=0$, which is false.
- Suppose $(p,r)=(\pm 1,0)$. Adjusting the sign of (p,q,r,s), we may assume that $(p,r)=(1,0)$. Then to be in $\mathrm{SL}_2(\mathbb{Z})$, we must have $s=1$ also, so by replacing r with m, we are merely looking at (3.3).

  As such, $a'=a$, and we claim that $b=b'$, which will be enough because $\operatorname{disc} f=\operatorname{disc} f'$. Well, $b'=b+2ma$, so $|b'|\le 2a$ forces b' except if $b\in\{\pm a\}$. But then $|b|=|b'|=a$, so being reduced forces $b=b'=a$.
- Suppose $(p,r)=(0,\pm 1)$. Again, we may adjust signs so that $(p,r)=(0,1)$; this then forces $q=-1$. Setting $s\coloneqq m$, we are now looking at

$$\begin{bmatrix} a' & b'/2 \\ b'/2 & c' \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & m \end{bmatrix} \begin{bmatrix} a & b/2 \\ b/2 & c \end{bmatrix} \begin{bmatrix} 0 & -1 \\ 1 & m \end{bmatrix} = \begin{bmatrix} c & mc-b/2 \\ mc-b/2 & a-mb+m^2c \end{bmatrix}.$$

  Now, the right-hand side needs to be a reduced form, so $|b-2mc|\le c$, which forces $m\in\{-1,0,1\}$ because $|b|\le c$ already. If $m=0$, then we are looking at $[a,b,c]\sim[c,-b,a]$, but then $a\le c\le a$ for both of these forms to be semi-reduced, meaning $a=c$, so $b,-b\ge 0$ for both of these forms to be reduced, meaning $b=0$, so $[a,b,c]=[a,0,a]=[c,-b,a]$.

  Otherwise, $m=\pm 1$. Note $c\ge|b\pm 2c|\ge 2c-|b|$, so $|b|\ge c\ge a\ge|b|$, so in fact $a=b=c$. But $a=c$ requires $b\ge 0$ in $[a,b,c]$ to be reduced, and similarly requires $-b\ge 0$ in $[c,-b,a]$ to be reduced, so again we conclude $[a,b,c]=[a,0,a]=[c,-b,a]$.
- Suppose $|p|=|q|=1$. Then (3.4) actually forces $a=a'$, but $a'=ap^2+bpr+cr^2=a\pm b+c$ then requires $b=\pm c$, so $c=|b|\le a\le c$ again. So equalities follow everywhere, but then $b\ge 0$ to be reduced, so $[a,b,c]=[a,a,a]$.

  Now, we may adjust signs so that $p=1$. Take $r=1$ for the moment. We then set $q\coloneqq m$ so that $s\coloneqq m+1$, from which we compute

$$\begin{bmatrix} a' & b'/2 \\ b'/2 & c' \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ m & m+1 \end{bmatrix} \begin{bmatrix} a & b/2 \\ b/2 & c \end{bmatrix} \begin{bmatrix} 1 & m \\ 1 & m+1 \end{bmatrix} = \begin{bmatrix} 3a & 3ma+3a/2 \\ 3ma+a/2 & \left(3m^2+3m+1\right)a \end{bmatrix}.$$

  Now, to be reduced, we need $|6ma+3a|\le 3a$, which requires $m\in\{0,-1\}$, but in either case we fail to have $3a\le\left(3m^2+3m+1\right)a$. On the other hand, if $r=-1$, we set $q\coloneqq -m$ so that $s\coloneqq m+1$, from which we compute

$$\begin{bmatrix} a' & b'/2 \\ b'/2 & c' \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ m & m+1 \end{bmatrix} \begin{bmatrix} a & b/2 \\ b/2 & c \end{bmatrix} \begin{bmatrix} 1 & m \\ 1 & m+1 \end{bmatrix} = \begin{bmatrix} a & -ma-a/2 \\ -ma-a/2 & \left(m^2+m+1\right)a \end{bmatrix}.$$

  Once again, we need $|-2ma-a|\le a$, which requires $m\in\{0,1\}$, but both cases then give $c'=\left(m^2+m+1\right)=a=a'$, so we are forced to have $b'=-2ma-a\ge 0$ to be reduced, meaning $b'=a$. So $[a,b,c]=[a,a,a]=[a',b',c']$. ∎

Observe that the proof of Theorem 3.30 is constructive: one can actually take a binary quadratic form $[a,b,c]$ and reduce it.

Example 3.31. We find a reduced binary quadratic form equivalent to $[4,5,2]$.

Solution. We use the moves (3.2) and (3.3) in succession, as described in the proof of Theorem 3.30. For brevity, let $[a,b,c]=[4,5,2]$.

1. Currently, $c<a$, so we use (3.2) and compute

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 4 & 5/2 \\ 5/2 & 2 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & -5/2 \\ -5/2 & 4 \end{bmatrix}.$$

2. Currently, $|b| > a$, so we use (3.2) with $m = 1$ and compute

$$\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & -5/2 \\ -5/2 & 4 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & -1/2 \\ -1/2 & 1 \end{bmatrix}.$$

3. Currently, $c < a$, so we use (3.2) and compute

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 2 & -1/2 \\ -1/2 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 2 \end{bmatrix}.$$

Our last step gives us the form $[1, 1, 2]$, which is indeed reduced. ■

3.1.4 Some Examples

Let's use Theorem 3.30 for fun and profit. For example, we can quickly reprove Example 3.10.

Lemma 3.32. Let D be a negative integer which is either 0 or $1 \pmod 4$. Then an odd integer n is properly represented by a reduced primitive binary quadratic form of discriminant D if and only if there is an integer b such that $b^2 \equiv D \pmod n$.

Proof. Corollary 3.26 implies that the conclusion is equivalent to being properly represented by some primitive binary quadratic form, and Proposition 3.24 allows us to replace any given primitive binary quadratic form by any equivalent one (which remains primitive by Remark 3.20), from which Theorem 3.30 allows us to assume that the form is reduced. ■

Thus, we are interested in computing reduced forms of a given discriminant. The following lemma will be helpful.

Lemma 3.33. Let D be a negative integer which is either 0 or $1 \pmod 4$. If $[a, b, c]$ is a reduced form of discriminant D, then $a \le \sqrt{-D/3}$, and $b \equiv D \pmod 2$.

Proof. For the bound on a, we note

$$-D = -b^2 + 4ac \ge -a^2 + 4a^2 = 3a^2,$$

so the inequality follows upon taking square roots. For the condition on b, note $D = b^2 - 4ac$ reduces $\pmod 2$ to $D \equiv b \pmod 2$. ■

Remark 3.34. Lemma 3.33 has the amazing consequence of telling us that the number of equivalence classes of binary quadratic forms of any given negative discriminant D (which is 0 or $1 \pmod 4$) will be finite. Indeed, Theorem 3.30 tells us to count reduced forms, so we are solving

$$-D = 4ac - b^2.$$

Now, Lemma 3.33 upper-bounds $a \le \sqrt{-D/3}$, but then $|b| \le a \le \sqrt{-D/3}$ has only finitely many options too, and once a and b are given $c = (b^2 - D)/(4a)$ is forced.

And here is another proof of Example 3.10, entirely avoiding Theorem 2.84.

Example 3.35. The form $[1, 1, 2]$ is the only reduced binary quadratic form of discriminant -7. Thus, for an odd prime p, there are integers (x, y) such that $p = x^2 + xy + 2y^2$ if and only if there is an integer b such that $b^2 \equiv -7 \pmod p$.

Proof. The second sentence follows from the first sentence via Lemma 3.32. So it remains to classify reduced binary quadratic forms $[a, b, c]$ of discriminant -7, for which we use Lemma 3.33 by doing casework on a: we want

$$a \le \sqrt{7/3} < 2,$$

where the second inequality is because $7 < 12 = (2\sqrt{3})^2$. Note $a = 0$ is impossible because we are looking for forms of negative discriminant, so we really only have $a = 1$ now. Then $|b| \le a = 1$ to be reduced, and Lemma 3.32 requires $b \equiv -7 \pmod 2$, so we have $b \in \{\pm 1\}$, but then $|b| = a$, so we must have $b \ge 0$, meaning $b = 1$. But we can now check that $(a, b) = 1$ enforces $c = \left(-7 - 1^2\right)/4 = 2$, so we are left with the reduced form $[1, 1, 2]$. ■

Exercise 3.36. Use the above method to give a new proof of Proposition 2.89.

Exercise 3.37. Use the above method to show that, for an odd prime p, there are integers (x, y) such that $p = x^2 + 2y^2$ if and only if there is an integer b such that $b^2 \equiv -2 \pmod p$.

For our troubles, we even get to improve Example 3.11.

Example 3.38. The form $[1, 1, 5]$ is the only reduced binary quadratic form of discriminant -19. Thus, for an odd prime p, there are integers (x, y) such that $p = x^2 + xy + 5y^2$ if and only if there is an integer b such that $b^2 \equiv -19 \pmod p$.

Solution. Again, the second sentence follows from the first via Lemma 3.32. It remains to classify binary quadratic forms $[a, b, c]$ of discriminant -19, for which we use Lemma 3.33 by doing casework on a; we want

$$a \le \sqrt{19/3} < 3,$$

where the second inequality is because $19 < 27$. Note $a = 0$ is impossible because we are looking forms of negative discriminant, so we have two cases to deal with. Note $b \equiv D \equiv 1 \pmod 2$ in all cases.

1. If $a = 1$, then $|b| \le 1$, but b is odd, so $|b| = 1 = a$, but to be reduced we then require $b \ge 0$, so $b = 1$. We can now check that $(a, b) = (1, 1)$ yields $c = \left(b^2 - D\right)/(4a) = (19 + 1)/4 = 5$, so we have $[1, 1, 5]$ in this case.
2. If $a = 2$, then $|b| \le 2$, but b is odd, so $b \in \{\pm 1\}$. In either case, $b^2 = 1$, so we must have $c = \left(b^2 - D\right)/(4a) = (19 + 1)/(4 \cdot 2)$, which is not an integer, so we get no reduced forms in this case.

Totaling our work, we see that $[1, 1, 5]$ is our only reduced form. ■

However, we are still unable to resolve Proposition 2.92 completely.

Example 3.39. There are two binary quadratic forms $[1, 0, 5]$ and $[2, 2, 3]$ of discriminant -20. Thus, for an odd prime p, there are integers (x, y) such that $p = x^2 + 5y^2$ or $p = 2x^2 + 2xy + 3y^2$ if and only if there is an integer b such that $b^2 \equiv -20 \pmod p$.

Solution. As usual, the second sentence follows from the first via Lemma 3.32, so we classify binary quadratic forms of discriminant -20 using Lemma 3.33. Our bound is

$$a \le \sqrt{20/3} < 3,$$

where the second inequality is because $20 < 27$. This time we want $b \equiv D \equiv 0 \pmod 2$ in all cases. As usual, we remove $a = 0$ because we want our discriminant to be negative.

1. If $a = 1$, then $|b| \leq 1$, but b is even, so $b = 0$. Then $c = \left(b^2 - D\right)/(4a) = 20/4 = 5$, so we have the reduced form $[1, 0, 5]$ in this case.

2. If $a = 2$, then $|b| \leq 2$, but b is even, so $b \in \{-2, 0, 2\}$. If $b = 0$, then $c = \left(b^2 - D\right)/(4a) = 20/8$ is not an integer, so we have no reduced form. Otherwise, $|b| = 2 = a$, so $b \geq 0$ to be reduced, so we have $b = 2$. This gives $c = \left(b^2 - D\right)/(4a) = 24/8 = 3$, so we have the reduced form $[2, 2, 3]$ in this case.

Totaling our work, we see that $[1, 0, 5]$ and $[2, 2, 3]$ are the only reduced forms in this case. ■

3.1.5 Quadratic Residues

Thus far we have been dealing with conditions like "for a prime p, there exists an integer b such that $b^2 \equiv -5 \pmod{p}$," but in practice, this appears to be a somewhat difficult condition to check, especially if p is large. For example, to check that it's false, it appears one would have to at least check all $b \in \{0, 1, \ldots, (p-1)/2\}$ to make sure nothing squares to $-5 \pmod{p}$. The goal of this subsection is to make it easier to check this result.

To begin, we will want to give language to equations of the type $b^2 \equiv -5 \pmod{p}$.

Definition 3.40 (quadratic residue)**.** Fix an odd prime p. Then an element $a \in (\mathbb{Z}/p\mathbb{Z})^\times$ is a *quadratic residue* if and only if there exists $b \in (\mathbb{Z}/p\mathbb{Z})^\times$ such that $a = b^2$; otherwise, we say that a is a *nonquadratic residue*. By convention, 0 is neither a quadratic residue nor a nonquadratic residue.

Example 3.41. Fix $p = 7$. Computing $(\pm 1)^2 \equiv 1$ and $(\pm 2)^2 \equiv 4$ and $(\pm 3)^2 \equiv 2 \pmod{2}$, we see that the quadratic residues $\pmod{7}$ are $\{1, 2, 4\}$, and the nonquadratic residues are $\{3, 5, 6\}$.

It appears that we are partitioning $(\mathbb{Z}/p\mathbb{Z})^\times$ into two pieces: squares and non-squares. Let's check that these classes are in fact the same size.

Lemma 3.42. Fix an odd prime p. Then there are exactly $(p-1)/2$ quadratic residues and $(p-1)/2$ nonquadratic residues. In fact, if $a \in (\mathbb{Z}/p\mathbb{Z})^\times$, then

$$a^{(p-1)/2} \equiv \begin{cases} +1 & \text{if } a \text{ is a quadratic residue,} \\ -1 & \text{if } a \text{ is a nonquadratic residue.} \end{cases}$$

Proof. The map $(\mathbb{Z}/p\mathbb{Z})^\times \to (\mathbb{Z}/p\mathbb{Z})^\times$ by $b \mapsto b^2$ is a surjection onto the set Q of quadratic residues. And for each $b^2 \in Q$, we note that the pre-image of b^2 consists of the $c \in (\mathbb{Z}/p\mathbb{Z})^\times$ such that $c^2 - b^2 \equiv 0 \pmod{p}$, which is equivalent to $(c+b)(c-b) \equiv 0 \pmod{p}$ or $c \equiv \pm b \pmod{p}$. Thus, the pre-image of each $b^2 \in Q$ has exactly two elements. Thus, $2 \cdot \#Q = p - 1$, so $\#Q = (p-1)/2$, so there are $(p-1)/2$ quadratic residues, leaving $(p-1)/2$ nonquadratic residues. ■

It will be useful to have an indicator of quadratic residues. For various reasons, the following indicator is best.

Definition 3.43 (Legendre symbol)**.** Fix a prime p. For $a \in \mathbb{Z}/p\mathbb{Z}$, we define the *Legendre symbol*

$$\left(\frac{a}{p}\right) := \begin{cases} +1 & \text{if } a \text{ is a quadratic residue,} \\ 0 & \text{if } a = 0, \\ -1 & \text{if } a \text{ is a nonquadratic residue.} \end{cases}$$

Approximately speaking, the following is why the Legendre symbol is a good indicator.

Proposition 3.44 (Euler's criterion)**.** Fix an odd prime p. For $a \in \mathbb{Z}/p\mathbb{Z}$, we have

$$\left(\frac{a}{p}\right) \equiv a^{(p-1)/2} \pmod{p}.$$

Proof. If $a = 0$, there is nothing to say, so we take $a \in (\mathbb{Z}/p\mathbb{Z})^\times$ in the argument which follows. Now, the main point is that

$$0 = a^{p-1} - 1 = \left(a^{(p-1)/2} - 1\right)\left(a^{(p-1)/2} + 1\right),$$

so at least $a^{(p-1)/2} \equiv \pm 1 \pmod{p}$. For example, if a is a quadratic residue so that $a \equiv b^2$, then $a^{(p-1)/2} \equiv b^{p-1} \equiv 1$, so it remains to show that $a^{(p-1)/2} \equiv -1 \pmod{p}$ if a is a nonquadratic residue.

In fact, we will show the contrapositive: suppose $a^{(p-1)/2} \equiv 1 \pmod{p}$, and we show that a is a quadratic residue. Considering the unique prime factorization of $x^{(p-1)/2} - 1$ in $\mathbb{F}_p[x]$, we see that this polynomial has at most $(p-1)/2$ linear factors in its prime factorization, and only linear factors are able to produce roots, so $x^{(p-1)/2} - 1$ has at most $(p-1)/2$ roots. But in light of Lemma 3.42, we have already given $(p-1)/2$ roots, so it follows that $a^{(p-1)/2} \equiv 1$ if and only if a is a quadratic residue. ■

Example 3.45. We re-prove Lemma 2.90. Fix an odd prime p. By Proposition 3.44, $-1 \pmod{p}$ is a square implies

$$(-1)^{(p-1)/2} \equiv \left(\frac{-1}{p}\right) \equiv 1 \pmod{p},$$

and in fact the reverse implication holds too because -1 not being a square would imply $(-1)^{(p-1)/2} \equiv -1 \pmod{p}$. But this is now equivalent to $p \equiv 1 \pmod{4}$.

Example 3.46. We show that 5 is not a quadratic residue of $p = 43$.

Solution. By Proposition 3.44 and the fact that $p \geq 3$, it is enough to check $5^{(p-1)/2} \equiv -1 \pmod{p}$. Thus, we are computing $5^{21} \pmod{43}$, for which we approximately do modular exponentiation by repeated squarings.

- Note $5^3 = 125 \equiv -4 \pmod{43}$.
- Note $5^9 \equiv (-4)^3 \equiv -64 \equiv -21 \pmod{43}$.
- Note $5^{18} \equiv (-21)^2 \cdot 441 \equiv 11 \pmod{43}$.
- Lastly, $5^{21} \equiv -4 \cdot 11 \equiv -44 \equiv -1 \pmod{43}$.

Our work has showed that $5^{21} \equiv -1 \pmod{43}$, so we are done. ■

Exercise 3.47. Show that 5 is a quadratic residue of $p = 43$.

At the very least, we have now reduced to the problem of checking that some $a \in (\mathbb{Z}/p\mathbb{Z})^\times$ is a square to taking exponents modulo primes, for which one can do by exponentiation by repeated squarings. However, we will be able to do better.

Remark 3.48. It is around this point that computing quadratic residues has become non-constructive: if Euler's criterion tells us that $a^{(p-1)/2} \equiv 1 \pmod{p}$, it is an entirely separate problem of actually finding a square root for a. (Recall that this is necessary to apply the algorithms suggested by section 3.1.4!) For some special cases, we refer to Problem 3.1.4 and Problem 3.1.7.

As an aside, we note that Proposition 3.44 has the corollary of showing that $\left(\frac{\cdot}{p}\right) : (\mathbb{Z}/p\mathbb{Z})^\times \to \{\pm 1\}$ is a group homomorphism. Note that this is consistent with Lemma 3.42.

Corollary 3.49. Fix an odd prime p. For any integers $a, b \in \mathbb{Z}$, we have

$$\left(\frac{a}{p}\right)\left(\frac{b}{p}\right) = \left(\frac{ab}{p}\right).$$

Proof. We have

$$\left(\frac{ab}{p}\right) \equiv (ab)^{p-1}/2 \equiv a^{(p-1)/2} \cdot b^{(p-1)/2} \equiv \left(\frac{a}{p}\right)\left(\frac{b}{p}\right) \pmod{p}.$$

But $p \geq 3$, so $\{-1, 0, 1\}$ are distinct $\pmod{p}$, so equality follows. ∎

3.1.6 Quadratic Reciprocity

Quadratic reciprocity will be an efficient tool for determining when some $a \in (\mathbb{Z}/p\mathbb{Z})^\times$ is a quadratic residue modulo an odd prime p. Because $\left(\frac{a}{p}\right)$ is multiplicative in a, it suffices to work when a is prime. Indeed, by unique prime factorization in $\mathbb{Z}$, one can factor a into

$$a = \varepsilon \prod_{i=1}^{n} q_i^{\nu_i}$$

where $\varepsilon \in \{\pm 1\}$, the $q_\bullet$ are distinct primes, and the ν_i are some positive integers. Then we may compute $\left(\frac{a}{p}\right)$, we see

$$\left(\frac{a}{p}\right) = \left(\frac{\varepsilon}{p}\right) \prod_{i=1}^{n} \left(\frac{q_i}{p}\right)^{\nu_i}.$$

We can compute $\left(\frac{\varepsilon}{p}\right)$ via Example 3.45, so it remains to compute $\left(\frac{q}{p}\right)$ where q is some prime. This is the content of quadratic reciprocity.

We will deal with the prime $q = 2$ separately. In some sense, this will be warm-up to our full proof of quadratic reciprocity.

Proposition 3.50. Fix an odd prime p. Then

$$\left(\frac{2}{p}\right) = \begin{cases} +1 & \text{if } p \equiv \pm 1 \pmod{8}, \\ -1 & \text{if } p \equiv \pm 3 \pmod{8}. \end{cases}$$

Proof. Consider $g_2 := \zeta_8 + \zeta_8^{-1}$, where $\zeta_8 = \exp(2\pi i/8) = \frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2}i$ is an eighth root of unity. The trick is to attempt to compute $g_2^p \pmod{p}$, where we are placing g_2 in the quotient ring $\mathbb{Z}[\zeta_8]/(p)$. We do this with two approaches.

1. The binomial theorem implies $(a+b)^p \equiv a^p + b^p$ in this ring, so we may write

   $$g_2^p = \left(\zeta_8 + \zeta_8^{-1}\right)^p \equiv \zeta_8^p + \zeta_8^{-p}.$$

   So if $p \equiv \pm 1 \pmod{8}$, then $g_2^p \equiv g_2$; if $p \equiv \pm 3 \pmod{8}$, then we have $\zeta_8^3 + \zeta_8^{-3} = -\zeta_8^{-1} - \zeta_8 = -g_2$.

2. We note that $\zeta_8 + \zeta_8^{-1} = \sqrt{2}$, so

   $$\left(\zeta_8 + \zeta_8^{-1}\right)^{p-1} = 2^{(p-1)/2} \equiv \left(\frac{2}{p}\right) \pmod{p},$$

   so $g_2^p \equiv \left(\frac{2}{p}\right) g_2 \pmod{p}$.

Combining the above cases, we see that

$$\left(\frac{2}{p}\right) g_2 \equiv g_2^p \equiv \begin{cases} +g_2 & \text{if } p \equiv \pm 1 \pmod 8, \\ -g_2 & \text{if } p \equiv \pm 3 \pmod 8. \end{cases}.$$

So we will be done as long as $g_2 \not\equiv 0 \pmod p$. Well, this would imply that $\zeta_8^4 \equiv 1 \pmod p$, so $-1 \equiv 1 \pmod p$, so $2 \equiv 0 \pmod p$, which is false because p is an odd prime. ■

Remark 3.51. There is actually some content to $2 \not\equiv 0 \pmod p$ because we are working in the ring $\mathbb{Z}[\zeta_8]/(p)$. Essentially, we are trying to show that $2 \notin p\mathbb{Z}[\zeta_8]$, which is equivalent to $2/p \notin \mathbb{Z}[\zeta_8]$. Well, $2/p \in \mathbb{Z}[\zeta_8]$ would imply that $2/p$ is an algebraic integer because ζ_8 is an algebraic integer because $\zeta_8^8 - 1 = 0$. (We are using Proposition 2.35 repeatedly.) But if $2/p$ is an algebraic integer, then Example 2.32 requires $2/p \in \mathbb{Z}$, which is false because p is an odd prime.

Example 3.52. Combining Exercise 3.37 with Proposition 3.50, we see that, for an odd prime p, there are integers (x, y) such that $p = x^2 + 2y^2$ if and only if $p \equiv \pm 1 \pmod 8$.

Let's now handle odd primes q.

Theorem 3.53 (quadratic reciprocity)**.** Fix distinct odd positive primes p and q. Then

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = (-1)^{\frac{p-1}{2}\cdot\frac{q-1}{2}}.$$

Theorem 3.53 is a truly amazing result: it tells us that determining when $x^2 \equiv p \pmod q$ somehow has to do with when $x^2 \equiv q \pmod p$, and the ability to switch like this makes the computation of these Legendre symbols quite efficient, akin to our computations in Proposition 1.11.

Before proving Theorem 3.53, let's see a quick example.

Example 3.54. Using Theorem 3.53 with Example 3.45, any odd prime p has

$$\left(\frac{-7}{p}\right) = \left(\frac{-1}{p}\right)\left(\frac{7}{p}\right) = (-1)^{\frac{p-1}{2}} \cdot (-1)^{\frac{p-1}{2}} \left(\frac{p}{7}\right) = \left(\frac{p}{7}\right).$$

Thus, plugging Example 3.41 into the above computation, Example 3.10 tells us that, for odd primes p, there are integers (x, y) such that $p = x^2 + xy + 2y^2$ if and only if $p \pmod 7 \in \{1, 2, 4\}$.

Thus, we no longer need to do a long and tedious check to see if $-7 \pmod p$ is a square: we simply check $p \pmod 7$ instead! This is an amazing improvement!

Anyway, let's prove Theorem 3.53.

Proof of Theorem 3.53. The proof is in the same general spirit as Proposition 3.50. Our main character will be the element

$$g_q := \sum_{k=1}^{q-1} \left(\frac{k}{q}\right) \zeta_q^k,$$

where $\zeta_q := \exp(2\pi i/q)$ is a qth root of unity. We will evaluate $g_q^p \pmod p$ two ways.

1. The binomial theorem still implies $(a+b)^p \equiv a^p + b^p$ in our ring $\mathbb{Z}[\zeta_q]/(p)$, so

$$g_q^p \equiv \sum_{k=1}^{q-1} \left(\frac{k}{q}\right) \zeta_q^{kp}.$$

Sending $k \in (\mathbb{Z}/q\mathbb{Z})^\times$ to kp^{-1}, we reparametrize the sum into

$$g_q^p \equiv \sum_{k=1}^{q-1} \left(\frac{k}{q}\right)\left(\frac{p}{q}\right)^{-1} \zeta_q^k = \left(\frac{p}{q}\right) g_q,$$

where we have used Corollary 3.49 and the fact that $\left(\frac{p}{q}\right) = \left(\frac{p}{q}\right)^{-1}$ in our simplification.

2. We show that $|g_q|^2 = q$. This is more or less a direct computation. We see

$$|g_q|^2 = g_q \cdot \overline{g_q} = \sum_{k,\ell=1}^{q-1} \left(\frac{k}{q}\right)\left(\frac{\ell}{q}\right)\zeta_q^{k-\ell}.$$

Now, the key to the computation is to replace ℓ with ak, letting a vary over $(\mathbb{Z}/q\mathbb{Z})^\times$. Then we have

$$|g_q|^2 = \sum_{a,k=1}^{q-1} \left(\frac{a}{q}\right) \underbrace{\left(\frac{k}{q}\right)^2}_{1} \zeta_q^{k(a-1)} = \sum_{a=1}^{q-1} \left(\left(\frac{a}{q}\right) \sum_{k=1}^{q-1} \zeta_q^{k(a-1)} \right).$$

To compute the inner sum, if $a = 1$, then we all terms are 1, so we have $q-1$. Otherwise, $\zeta_q^{a-1} \neq 1$, so the inner sum is a geometric sum with common ratio ζ_q^{a-1}, so we may evaluate it as

$$\sum_{k=1}^{q-1} \zeta_q^{k(a-1)} = -1 + \sum_{k=0}^{q-1} \zeta_q^{k(a-1)} = -1 + \frac{\zeta_q^{q(a-1)} - 1}{\zeta_q^{a-1} - 1} = -1.$$

Totaling, we see

$$|g_q|^2 = (q-1) - \sum_{a=2}^{q-1} \left(\frac{a}{q}\right) = q - \sum_{a=1}^{q-1} \left(\frac{a}{q}\right).$$

Now, Lemma 3.42 tells us that half of $(\mathbb{Z}/q\mathbb{Z})^\times$ is a quadratic residue, and the other half is a non-quadratic residue, so the current summation evaluates to $\frac{q-1}{2} - \frac{q-1}{2} = 0$, finishing.

We now complete the proof. For the second computation, we note

$$\overline{g_q} = \sum_{k=1}^{q-1} \left(\frac{k}{q}\right)\zeta_q^{-k} = \sum_{k=1}^{q-1} \left(\frac{-k}{q}\right)\zeta_q^{k} = \left(\frac{-1}{q}\right) g_q,$$

so

$$q = |g_q|^2 = g_q \cdot \overline{g_q} = \left(\frac{-1}{q}\right) g_q^2,$$

so synthesizing our casework reveals

$$\left(\frac{p}{q}\right) g_q \equiv g_q^p = \left(g_q^2\right)^{(p-1)/2} g_q = \left(\left(\frac{-1}{q}\right) q\right)^{(p-1)/2} g_q \equiv (-1)^{\frac{p-1}{2}\cdot\frac{q-1}{2}} \left(\frac{q}{p}\right) g_q,$$

where we have used Proposition 3.44 in the last step, so

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) g_q \equiv (-1)^{\frac{p-1}{2}\cdot\frac{q-1}{2}} g_q \pmod p.$$

Now, $|g_q|^2 = q$ is nonzero $\pmod p$, so we may cancel it on both sides, completing the proof. If we wish to be as rigorous as in Remark 3.51, we can multiply both sides above by $\overline{g_q}$ to see that

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) q \equiv (-1)^{\frac{p-1}{2}\cdot\frac{q-1}{2}} q \pmod p,$$

but then q has an inverse $\pmod p$ (in $\mathbb{Z}$), so we may multiply both sides by this to see that

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) \equiv (-1)^{\frac{p-1}{2}\cdot\frac{q-1}{2}} \pmod p.$$

At this point, everything is a sign, so to show that we must have the same sign on both sides, it is enough to show that $1 \not\equiv -1 \pmod p$, or $2 \not\equiv 0 \pmod p$, which is exactly Remark 3.51. ■

Let's see another example to wrap ourselves up, finally resolving Proposition 2.92.

Example 3.55. Fix an odd prime p. Then there are integers (x, y) such that $p = x^2 + 5y^2$ if and only if $p \pmod{20} \in \{1, 9\}$.

Proof. Looking at Example 3.39, we first want to determine when $\left(\frac{-20}{p}\right) = 1$. By Corollary 3.49, it is equivalent to compute $\left(\frac{-5}{p}\right)$, for which we use Theorem 3.53 with Example 3.45 to see

$$\left(\frac{-5}{p}\right) = \left(\frac{-1}{p}\right)\left(\frac{5}{p}\right) = (-1)^{(p-1)/2}\left(\frac{p}{5}\right).$$

Now, the quadratic residues $\pmod 5$ are $\{(\pm1)^2, (\pm2)^2\} = \{1, 4\}$, so we see

$$\left(\frac{-5}{p}\right) = \begin{cases} +1 & \text{if } p \pmod{20} \in \{1, 3, 7, 9\}, \\ -1 & \text{if } p \pmod{20} \in \{11, 13, 17, 19\}. \end{cases}$$

Thus, Example 3.39 tells us that $p \pmod{20} \in \{1, 3, 7, 9\}$ is equivalent to having integers (x, y) such that $p = x^2 + 5y^2$ or $p = 2x^2 + 2xy + 3y^2$.

However, from the available options, a direct computation shows that $x^2 + 5y^2$ is only ever in $1 \pmod{20}$ or $9 \pmod{20}$ or $7 \pmod{20}$ by a direct computation; similarly, one checks $2x^2 + 2xy + 3y^2 \pmod{20}$ is only ever in $3 \pmod{20}$. Thus, for $p \pmod{20} \in \{1, 9\}$, we must have $p = x^2 + 5y^2$; and conversely, the other cases $p \pmod{20} \in \{3, 7\}$ cannot have $p = x^2 + 5y^2$. ■

Remark 3.56. The last paragraph about to distinguish $x^2 + 5y^2$ from $2x^2 + 2xy + 3y^2$ is the beginning of "genus theory," a topic which is sadly just barely beyond the scope of this course.

3.1.7 Problems

Do ten points worth of the following exercises.

Problem 3.1.1 (1 point)**.** Find the unique reduced binary quadratic form which is equivalent to $[10, 15, 6]$.

Problem 3.1.2 (1 point)**.** Compute the Legendre symbol

$$\left(\frac{61}{107}\right).$$

Problem 3.1.3 (2 points)**.** Show that a binary quadratic form f is primitive if and only if

$$\gcd_{(x,y)\in\mathbb{Z}^2} f(x, y) = 1.$$

Problem 3.1.4 (2 points). Let $p \equiv 3 \pmod 4$ be a prime. For any $x \in (\mathbb{Z}/p\mathbb{Z})^\times$ which is a quadratic residue, show that $y := x^{(p+1)/4}$ has $y^2 \equiv x \pmod p$. In general, it is a little difficult to compute square roots modulo primes.

Problem 3.1.5 (3 points). For an odd prime p, there are integers (x, y) such that $p = x^2 + xy + 3y^2$ if and only if there is an integer b such that $b^2 \equiv -11 \pmod p$. Describe all such p via congruence conditions $\pmod{11}$.

Problem 3.1.6 (4 points). Given a negative integer D which is either 0 or $1 \pmod 4$, write a computer program outputting all reduced binary quadratic forms of discriminant D in the form $[a, b, c]$.

Problem 3.1.7 (7 points). Fix a prime $p \equiv 1 \pmod 4$.

(a) (1 point) If $x \in (\mathbb{Z}/p\mathbb{Z})^\times$ is not a quadratic residue, show that $y := x^{(p-1)/4}$ has $y^2 \equiv -1 \pmod p$.

(b) (2 points) Use (a) to write a computer program with probabilistically good running time which takes in a prime $p \equiv 1 \pmod 4$ and outputs some integer x such that $x^2 \equiv -1 \pmod p$.

(c) (4 points) Use (b) to write a computer program with probabilistically good running time which takes in a prime $p \equiv 1 \pmod 4$ and outputs a pair of integers (a, b) such that $p = a^2 + b^2$. Use the program to write the prime number $p = 10^{30} + 57$ as the sum of two squares.

3.2 (Almost) Unique Factorization

3.2.1 The Class Group

3.2.2 Class Groups of Imaginary Quadratic Fields

3.2.3 Back to Diophantine Equations

Theme 4
Intermission: Localization

That something so small could be so beautiful.

—Anthony Doerr, [Doe14]

4.1 Local Fields

In this section, we introduce the main characters of our intermission, which are the characteristic-0 local fields.

4.1.1 The p-adics, Algebraically

Fix a prime p in the following discussion. The goal of the present subsection is to define the ring of p-adic integers $\mathbb{Z}_p$. Approximately speaking, the point will be that $\mathbb{Z}_p$ is able to capture the information of $\mathbb{Z}/p\mathbb{Z}$ and $\mathbb{Z}/p^2\mathbb{Z}$ and $\mathbb{Z}/p^3\mathbb{Z}$ and so on, all at once.

To see why this might be helpful, we recall the idea of "local obstructions" we saw much earlier in this course: an equation might have solutions $\mathbb{Z}/p\mathbb{Z}$ but lose solutions in $\mathbb{Z}/p^2\mathbb{Z}$, or it might have solutions in $\mathbb{Z}/p^3\mathbb{Z}$ but lose solutions in $\mathbb{Z}/p^4\mathbb{Z}$.

Example 4.1. The equation $x^2 + 1 = 0$ has solutions in $\mathbb{Z}/2\mathbb{Z}$ but no solutions in $\mathbb{Z}/4\mathbb{Z}$.

Example 4.2. The equation $4x^2 + 4 = 0$ has solutions in $\mathbb{Z}/8\mathbb{Z}$ but no solutions in $\mathbb{Z}/16\mathbb{Z}$.

The ring $\mathbb{Z}_p$ will be able to keep track of all this modular information at once. For example, thinking about $\mathbb{Z}_p$ will allow us to cleanly thing about statements such as the following one.

Proposition 4.3. Let a be an integer. The following are equivalent.

(a) The equation $x^2 \equiv a \pmod 8$ has solutions.

(b) The equation $x^2 \equiv a \pmod{2^\nu}$ for any power 2^ν.

With this in mind, a good first guess for $\mathbb{Z}_p$ would be the infinite product ring

$$R_p \coloneqq \prod_{\nu=0}^{\infty} \mathbb{Z}/p^\nu\mathbb{Z}.$$

However, this ring has many problems. For example, R_p is not an integral domain: $(1,0,0,\dots)\cdot(0,1,0,\dots) = (0,0,0,\dots)$. A worse problem is that it somehow fails to actually care about modular information: a central property of the $\mathbb{Z}/p^\nu\mathbb{Z}$s is that if an equation has solutions in $\mathbb{Z}/p^{\nu+1}\mathbb{Z}$, then it will have solutions in $\mathbb{Z}/p^\nu\mathbb{Z}$ because of the reduction map

$$\mathbb{Z}/p^{\nu+1}\mathbb{Z} \to \mathbb{Z}/p^\nu\mathbb{Z}. \tag{4.1}$$

This reduction map would allow us to recover $\mathbb{Z}/p^\nu\mathbb{Z}$-solutions from "higher-order" solutions, and we would like $\mathbb{Z}_p$ to mirror this structure.

In particular, we would like projection maps $\mathbb{Z}_p \to \mathbb{Z}/p^\nu\mathbb{Z}$ for each p^ν, and we want these maps to commute with the maps (4.1). This allows us to fix R_p into the following ring.

Definition 4.4 (p-adic integers)**.** Fix a prime p. Then we define the ring of *p-adic integers* by

$$\mathbb{Z}_p \coloneqq \left\{ (a_\nu)_{\nu=0}^{\infty} \in \prod_{\nu=0}^{\infty} \mathbb{Z}/p^\nu\mathbb{Z} : a_{\nu+1} \equiv a_\nu \pmod{p^\nu} \right\}.$$

In practice, it is helpful to view $(a_\nu)_\nu$ as a sequence of integers, but it is important to remember that we only care about $a_\nu \pmod{p^\nu}$.

Here are some basic checks on $\mathbb{Z}_p$.

Lemma 4.5. Fix a prime p. Then $\mathbb{Z}_p$ is a subring of R_p.

Proof. We have the following checks.

- Identities: note that $(0,0,\dots) \in \mathbb{Z}_p$ and $(1,1,\dots) \in \mathbb{Z}_p$ because $a \equiv a \pmod{p^\nu}$ for any p^ν and $a \in \{0,1\}$.
- Closure: given $(a_\nu)_\nu, (b_\nu)_\nu \in \mathbb{Z}_p$, we note that

$$\begin{aligned} a_{\nu+1} + b_{\nu+1} &\equiv a_\nu + b_\nu \pmod{p^\nu}, \\ a_{\nu+1} \cdot b_{\nu+1} &\equiv a_\nu \cdot b_\nu \pmod{p^\nu}, \end{aligned}$$

  so $(a_\nu + b_\nu)_\nu, (a_\nu + b_\nu) \in \mathbb{Z}_p$ as well. ∎

We complained that R_p is not an integral domain, but $\mathbb{Z}_p$ is.

Lemma 4.6. Fix a prime p. Then $\mathbb{Z}_p$ is an integral domain.

Proof. The point is to use the coherence of the coordinates of in elements of $\mathbb{Z}_p$. Suppose we have nonzero $(a_\nu)_\nu, (b_\nu)_\nu \in \mathbb{Z}_p$, and we show that the product is nonzero. Because $(a_\nu)_\nu$ and $(b_\nu)_\nu$ are nonzero, we can find some specific nonnegative integers m and n such that $a_m \not\equiv 0 \pmod{p^m}$ and $b_n \not\equiv 0 \pmod{p^n}$.

Thus, the largest power of p dividing a_m is less than m, and the same then holds for a_{m+n}; similarly, and the largest power of p dividing b_n is less than n, and the same then holds for b_{m+n}. Thus, the largest power of p dividing $a_{m+n}b_{m+n}$ is less than $m+n$, so

$$a_{m+n}b_{m+n} \not\equiv 0 \pmod{p^{m+n}},$$

so $(a_\nu)_\nu \cdot (b_\nu)_\nu \neq 0$. ∎

It is notable that there is a copy of $\mathbb{Z}$ in $\mathbb{Z}_p$, and $\mathbb{Z}$ is "dense" in $\mathbb{Z}_p$ in some sense. This tells us that $\mathbb{Z}_p$ is indeed a fairly good approximation of $\mathbb{Z}$, but do remember that $\mathbb{Z}_p$ only cares about $\pmod{p^\nu}$ information.

4.1.2 Hensel's Lemma

4.1.3 Ostrowski's Theorem

4.1.4 Problems

Problem 4.1.1. Show that $x^2 - 223y^2 = -3$ has a solution in $\mathbb{Z}_3$ by showing that there exists $y \in \mathbb{Z}_3$ such that $y^2 = 4/223$. Use Problem 2.1.4 to conclude that the equation

$$x^2 - 223y^2 \equiv -3 \pmod{n}$$

has a solution for any positive integer n.

Theme 5
Cubic Equations

Every person believes that he knows what a curve is until he has learned so much mathematics that the countless possible abnormalities confuse him.

—Felix Klein, [Kle16]

Appendix A
Some Algebra

There was nothing clever to say, so I said something foolish

—Madeline Miller

A.1 Unique Factorization Domains

The goal of the present section is to review the notion of a unique factorization domain and basic properties of them. This is a notion we will want later in section 3.2, though it is not clear why yet.

Definition A.1 (integral domain)**.** A ring A is an *integral domain* if and only if $a \cdot b = 0$ implies that $a = 0$ or $b = 0$.

Example A.2. The ring $\mathbb{Z}$ is an integral domain. Any field is an integral domain.

The best integral domains are unique factorization domains. To recall this definition, we need the notion of prime and irreducible elements.

Definition A.3 (prime)**.** Fix a ring A. Then an element $p \in A$ is *prime* if and only if p is nonzero and the ideal (p) of A is a prime ideal. In other words, p is not zero, not a unit, and whenever $p \mid ab$ for $a, b \in A$, we have $p \mid a$ or $p \mid b$.

Definition A.4 (irreducible)**.** Fix a ring A. Then an element $p \in A$ is *irreducible* if and only if any factorization $p = ab$ has exactly one of a or b equal to a unit. Notably, p cannot be zero (for we could set $a = b = 0$), and p cannot be a unit (for then both a and b would be a unit).

Definition A.5 (unique factorization domain)**.** Fix an integral domain A. Then A is a *unique factorization domain* if and only if any nonzero element $r \in A$ has a unique factorization into irreducibles

$$r = \prod_{i=1}^{n} p_i,$$

where the sequence $\{p_i\}_{i=1}^{n}$ of irreducible elements of A is unique up multiplication by a unit and permutation.

An elementary number theory course would show that an integer is prime if and only if it is irreducible and then deduce that $\mathbb{Z}$ is a unique factorization domain. An algebra course would show the more general result that a principal ideal domain is a unique factorization domain. We will quickly review these arguments, but we will not dwell on them.

Proposition A.6. The ring $\mathbb{Z}$ is a principal ideal domain.

Proof. Let $I \subseteq \mathbb{Z}$ be an ideal. If $I = \{0\}$, then $I = (0)$. Otherwise, I contains a nonzero element $n \in I$, so I contains a positive element $n^2 \in I$. Thus, we may let $g \in I$ denote the least positive element. We claim that $I = (g)$; certainly $(g) \subseteq I$, so we want to show $I \subseteq (g)$. Well, choose any $a \in I$. Then we may use division to find integers $q, r \in \mathbb{Z}$ such that

$$a = gq + r$$

where $0 \le r < g$. Thus, $r = a - gq \in I$ is a nonnegative element of I strictly less than g, so r cannot be positive, so $r = 0$, so $a = gq$, so $a \in (g)$. ∎

We now move towards showing that principal ideal domains are unique factorization domains.

Lemma A.7. Fix an integral domain A. If p is prime, then p is irreducible.

Proof. Note that p is neither zero nor a unit by hypothesis. Suppose we factor $p = ab$ where $a, b \in A$; certainly both a and b cannot both units because then p would be a unit, so it remains to show that one is. Then $p \mid ab$, so $p \mid a$ or $p \mid b$ because p is prime. Without loss of generality, take $p \mid a$, and write $a = pa'$ so that

$$p = pa'b.$$

Then $1 = a'b$, so b is a unit. ∎

Lemma A.8. Let A be a principal ideal domain. If $p \in A$ is irreducible, and if $a \in A$ lives outside (p), then $(a, p) = A$. In other words, (p) is a maximal ideal.

Proof. Note that the last sentence follows from the previous because (p) would then be proper ideal with no ideal between (p) and A, making (p) maximal.

Now, note (a, p) is a principal ideal, so say $(a, p) = (d)$. However, because d divides p, we may write $p = de$ where e is an integer. Thus, one of d or e is a unit. We claim that d is a unit, which will complete the proof. Well, if e is a unit, then $d = pe^{-1}$, so pe^{-1} divides a, so p divides a (recall e is a unit), which is a contradiction. ∎

Remark A.9. One can interpret Lemma A.8 as showing that $A/(p)$ is a field for any irreducible p. Indeed, this follows from (p) being a maximal ideal.

Proposition A.10. Let A be a principal ideal domain. Then any irreducible element $p \in A$ is prime.

Proof. Note that p is neither zero nor a unit by hypothesis. Now, suppose we have $p \mid ab$ but $p \nmid a$. We want to show that $p \mid b$. Well, Lemma A.8 tells us that $(a, p) = A$, so we can write

$$ar + ps = 1$$

for some integers $r, s \in A$, so we see that

$$abr + psb = b,$$

so $p \mid ab$ and $p \mid psb$ implies that $p \mid b$, which is what we wanted. ∎

Theorem A.11. Any principal ideal domain A is a unique factorization domain.

We will split the proof of Theorem A.11 into a few parts. To begin, we prove existence, which does not require Proposition A.10.

Lemma A.12. Fix an integral domain A. Suppose that any ascending chain of principal ideals

$$(a_0) \subseteq (a_1) \subseteq (a_2) \subseteq \cdots$$

eventually stabilizes; in other words, there is some nonnegative integer N such that $(a_n) = (a_N)$ for any $n \geq N$. Then every nonzero element of A has a factorization into irreducibles.

Proof. Units take the "empty" factorization consisting of only the unit itself and no irreducibles. Irreducible elements attain the factorization consisting of the irreducible itself.

Now, suppose for the sake of contradiction that we have a nonzero element $r_0 \in A$ with no factorization into irreducibles. The work above shows that r_0 is not a unit and not irreducible, so it follows that we can factor $r_0 = s_1 r_1$ where neither s_1 nor r_1 is a unit or zero. If s_1 and r_1 both had factorizations into irreducibles, then we could multiply the factorizations together to produce a factorization for r_0. Thus, at least one of s_1 or r_1 cannot have a factorization into irreducibles; without loss of generality, it is r_1.

Iterating the process of the previous paragraph produces a sequence of elements $\{r_n\}_{n=0}^{\infty}$ and $\{s_n\}_{n=1}^{\infty}$ such that $r_n = r_{n+1} s_{n+1}$ for each n. But then we have the descending chain of principal ideals

$$(r_0) \subseteq (r_1) \subseteq (r_2) \subseteq \cdots,$$

which must stabilize eventually. Thus, there is some n for which $(r_n) = (r_{n+1})$, so we may find s' such that $r_{n+1} = s' r_n$ and thus

$$r_n = s_n r_{n+1} = s_n s' r_n.$$

This implies that $s_n s' = 1$ and so s_n is a unit, which is a contradiction to its construction. ∎

Remark A.13. More generally, a ring A will be called "Noetherian" if and only if any ascending chain of ideals stabilizes. We will avoid using this notion in these notes when possible, largely because it is not strictly necessary for the story we wish to tell.

Lemma A.14. Fix an integral domain A. Suppose that an element $p \in A$ is prime if and only if it is irreducible. Then for any equal factorizations of irreducibles

$$\prod_{i=1}^{m} p_i = \prod_{j=1}^{n} q_j,$$

we must have $m = n$, and there is a permutation σ of $\{1, 2, \ldots, n\}$ such that p_i and $q_{\sigma(i)}$ are the same up to multiplication by a unit.

Proof. Fix a factorization as hypothesized. We will induct on m. If $m = 0$, then all the $q_\bullet$ multiply out to 1 and hence by units, which makes no sense if $n > 0$. As such, the right-hand side must also be empty, meaning that $n = 0$, so there is nothing to prove. Note that a symmetric argument deduces that $n = 0$ implies $m = 0$, so we may assume that $m, n > 0$ in the argument which follows.

Now, for the induction, the hypothesis tells us that the irreducible p_m is prime and therefore must divide some factor $q_\bullet$ on the right-hand side. Adjusting via a permutation, we may assume that $p_m \mid q_n$. Then we may write $q_n = up_m$ for some $u \in A$, but in fact because p_m fails to be a unit, we see u is a unit because q_n is irreducible. As such, adjusting by a unit, we may assume that $q_n = p_m$, whereupon dividing both of our factorizations out by this redundancy leaves us with

$$\prod_{i=1}^{m-1} p_i = \prod_{j=1}^{n-1} q_j,$$

and now we may induct downwards. ∎

In fact, one has the following converse to Lemma A.14.

Lemma A.15. If A is a unique factorization domain, then an element $p \in A$ is prime if and only if it is irreducible.

Proof. The forward direction is covered by Lemma A.7. For the converse, suppose p is irreducible. Certainly p is not zero and not a unit. Now, suppose $p \mid ab$ for some $a, b \in A$. If $a = 0$, then $p \mid a$; similar holds if $b = 0$. Otherwise, we may give a and b factorizations into irreducibles by

$$a = \prod_{i=1}^{m} p_i \qquad \text{and} \qquad b = \prod_{j=1}^{n} q_j.$$

Because p divides the product ab, we see that $ab/p \in A$ will also have a factorization

$$\prod_{k=1}^{s} r_k = \frac{ab}{p},$$

so

$$p \prod_{k=1}^{s} r_k = ab = \prod_{i=1}^{m} p_i \cdot \prod_{j=1}^{n} q_j.$$

By the uniqueness of our factorizations, we see that the irreducible p (perhaps times a unit) must appear as one of the $p_\bullet$ or $q_\bullet$, so $p \mid a$ or $p \mid b$. ∎

We are now ready to prove Theorem A.11.

Proof of Theorem A.11. By Lemmas A.12 and A.14, it suffices to do the following two checks.

- Suppose we have an ascending chain of principal ideals

  $$(a_0) \subseteq (a_1) \subseteq (a_2) \subseteq \cdots,$$

  and we want to show that it stabilizes. Well, the union

  $$I \coloneqq \bigcup_{i=0}^{\infty} (a_i) = (a_0, a_1, a_2, \ldots)$$

  is an ideal of A. But all ideals are principal, so we may write $I = (a)$ for some $a \in A$. But then $a \in (a_N)$ for some N, so $I \subseteq (a_N)$, so for any $n \geq N$, we see that

  $$(a_n) \subseteq I \subseteq (a_N) \subseteq (a_n),$$

  establishing that our chain of ideals has stabilized.
- Note that an element of A is prime if and only if it is irreducible by combining Lemma A.7 and proposition A.10. ∎

A.2 A Little Field Theory

The notes assume ring and group theory, but we will spend this appendix establishing the field theory that we will need.

A.2.1 Basic Notions

Here is our following definition.

Definition A.16 (field)**.** A *field* K is a ring where each nonzero element has a multiplicative inverse.

We will be interested in how fields relate to each other.

Definition A.17 (field extension)**.** A *field extension* L/K is when one field K is contained in another L. The *degree* $[L : K]$ of the extension is the dimension of L as a K-vector space. The field extension is said to be *finite* if and only if $[L : K] < \infty$.

Example A.18. Fix a field extension L/K. Given $\alpha \in L$, we can construct the field $K(\alpha)$ which is the quotient field of the integral domain generated by K and $\alpha \in L$. Formally, $K(\alpha)$ is the quotient field of the ring $K[\alpha]$ which is defined as the image of the ring homomorphism

$$\operatorname{ev}_\alpha \colon K[x] \to L$$

defined by sending the polynomial $f(x) \in K[x]$ to $f(\alpha)$.

It might feel a little weird that we have jumped directly to one field being contained in another instead of a tamer notion of homomorphism. The following result explains why.

Lemma A.19. Let $\varphi \colon K \to L$ be a ring homomorphism of fields. Then φ is injective.

Proof. It suffices to check that $\ker \varphi$ is trivial. Well, $\ker \varphi$ is an ideal of K, but if nontrivial, then $\ker \varphi$ contains a unit and therefore must be all of K. However, $\varphi(1) = 1$, so $\ker \varphi \neq K$, so we see that we must have $\ker \varphi = 0$. ∎

A basic fact about our extensions is how degrees behave in extensions.

Lemma A.20. Let M/L and L/K be extensions of fields. Then

$$[M : L][L : K] = [M : K].$$

Proof. If M/L or L/K are infinite, then the K-vector space M will have an infinitely linearly independent set, meaning $[M : K]$ is also infinite. Otherwise, we take $[M : L] = [L : K]$ to be finite. Let $\{m_1, \dots, m_r\}$ and $\{\ell_1, \dots, \ell_s\}$ be bases for M/L and L/K, respectively. We claim that

$$\{m_i \ell_j\}_{1 \le i \le r, 1 \le j \le s}$$

is a basis for M/K.

- We show that the $m_i \ell_j$ span: any $m \in M$ can be expressed as

$$m = \sum_{i=1}^{r} a_i m_i$$

where $a_k \in L$, but then each $a_k \in L$ can be expressed as

$$m = \sum_{i=1}^{r} \sum_{j=1}^{s} b_{ij} m_i \ell_j,$$

where $b_{ij} \in K$. This is what we wanted.

- We show that the $m_i \ell_j$ are linearly independent: suppose

$$\sum_{i=1}^{r} \sum_{j=1}^{s} b_{ij} \ell_j m_i = \sum_{i=1}^{r} \sum_{j=1}^{s} b'_{ij} \ell_j m_i.$$

Then

$$\sum_{i=1}^{r} \sum_{j=1}^{s} (b_{ij} - b'_{ij}) \ell_j m_i = 0,$$

so because $\{m_1, \dots, m_r\}$ is a basis of M/L, we see

$$\sum_{j=1}^{s} (b_{ij} - b'_{ij}) \ell_j = 0$$

for each i, but because $\{\ell_1, \dots, \ell_s\}$ is a basis of L/K, we see $b_{ij} = b'_{ij}$ for each i and j. ■

A.2.2 Polynomial Rings

In this subsection, we show that $K[x]$ is a unique factorization domain for any field K. We will not bother to show the usual facts about degree over integral domains, such as $\deg fg = \deg f + \deg g$ for nonzero $f, g \in K[x]$. However, we will show the following result, whose proof is more technical than one would like. Thankfully, we will not have to use this result for a while.

Lemma A.21. Fix an irreducible polynomial $f \in \mathbb{C}[x]$. Then f has no repeated roots.

Proof. Note $\deg f \geq 1$ because $\deg f = 1$ implies that f is a unit and thus not irreducible. Because $\mathbb{C}$ is algebraically closed, we note that $f(x)$ factors as

$$f(x) = c \prod_{i=1}^{n} (x - \alpha_i)$$

for some complex numbers $c, \alpha_1, \dots, \alpha_n \in \mathbb{C}$. Now, if $\alpha_i = \alpha_j$ for $i \neq j$, then $f(x)$ has a double root at α_i, so a direct computation shows that $f'(\alpha_i) = 0$, so $f(x)$ and $f'(x)$ have a root in common, so $\gcd(f(x), f'(x))$ is a non-constant polynomial with degree strictly less than $f(x)$ but dividing $f(x)$, which contradicts $f(x)$ being irreducible. ■

Anyway, the main point to showing that $K[x]$ is a unique factorization domain is the following result.

Proposition A.22 (division)**.** Fix a field K, and let $a, b \in K[x]$ be polynomials with $b \neq 0$. Then there exist polynomials $q, r \in K[x]$ such that

$$a = bq + r$$

where $r = 0$ or $0 \leq \deg r < \deg b$.

Proof. We induct on $\deg a$. If $a = 0$ or $\deg a < \deg b$, we set $q = 0$ and $r = a$. Otherwise, $\deg a \geq \deg b$. Let the leading coefficient of b be $b_n x^n$, and let the leading coefficient of a be $a_m x^m$. Then

$$c(x) := a(x) - \frac{a_m}{b_n} x^{m-n} \cdot b$$

cancels out the leading coefficient of a, so $c = 0$ or $\deg c < \deg a$. So we can apply the result to c by the induction, writing

$$c = bq_c + r_c,$$

so

$$a(x) = b(x)\left(q_c(x) + \frac{a_m}{b_n} x^{m-n}\right) + r(x),$$

finishing. ■

Theorem A.23. Fix a field K. Then the field $K[x]$ is a principal ideal domain and hence a unique factorization domain.

Proof. If we show that $K[x]$ is a principal ideal domain, we finish immediately by Theorem A.11. So we want to show that $K[x]$ is a principal ideal domain.

Well, let $I \subseteq K[x]$ be a principal ideal domain. If $I = \{0\}$, then $I = (0)$, so there is nothing to say. Otherwise, I has a nonzero element, so we let $f \in I$ denote any element of least degree. We claim $I = (f)$. Certainly $(f) \subseteq I$, so we want to show that $a \in I$ lives in (f). Well, by Proposition A.22, we may write

$$a = fq + r$$

where $r = 0$ or $0 \leq \deg r < \deg f$. However, $r = a - fq \in I$ has $\deg r < \deg f$, so minimality of $\deg f$ requires $r = 0$, meaning $a = fq$, so $a \in (f)$. ■

A.2.3 Algebraic Elements

In this subsection, we show some basic properties of algebraic elements. Here is our definition.

Definition A.24 (algebraic)**.** Fix a field extension L/K. An element $\alpha \in L$ is *algebraic over* K if and only if α is the root of some nonzero polynomial in $K[x]$.

Example A.25. Any element of K is algebraic over K because $\alpha \in K$ is the root of the polynomial $x - \alpha \in K[x]$.

Finite extensions provide a wealth of algebraic elements.

Lemma A.26. Let L/K be a finite extension of fields. Then each $\alpha \in L$ is algebraic over K.

Proof. The elements $1, \alpha, \alpha^2, \ldots$ form an infinite set in L, so they cannot be K-linearly independent because $\dim_K L < \infty$. Thus, there is a relation of the form

$$\sum_{k=0}^{n} a_k \alpha^k = 0$$

where $a_k \in K$ are not all zero. As such, the polynomial $f(x) := \sum_{k=0}^{n} a_k x^k$ will do. ■

It will shortly be helpful to limit the polynomial attached to α somewhat.

Lemma A.27. Fix a field extension L/K, and let $\alpha \in L$ be algebraic over K. Then α is the root of a unique monic irreducible polynomial $f(x) \in K[x]$. In fact, for any polynomial $g \in K[x]$ with $g(\alpha) = 0$, we have $f \mid g$.

Proof. We begin by showing existence. We know that α is the root of some nonzero polynomial $f(x) \in K[x]$, so we choose $f(x)$ to have the smallest degree possible. By dividing out the leading coefficient (which is nonzero because f is nonzero), we may assume that f is monic. It remains to show that f is irreducible. Well, suppose that

$$f = ab$$

for $a, b \in K[x]$. Note neither a nor b is zero because this would imply $f = 0$; additionally, if both are units, then f is a unit and hence a constant polynomial, which also makes no sense. Now, evaluating at α, we see that $f(\alpha) = 0$ requires $a(\alpha) = 0$ or $b(\alpha) = 0$, so by minimality of f, we must have $\deg a \geq \deg f$ or $\deg b \geq \deg f$. Without loss of generality take $\deg a \geq \deg f$, but $f = ab$ then forces $\deg a = \deg f$, and b is a constant polynomial.

We now show that $g(\alpha) = 0$ implies $f \mid g$ for any $g \in K[x]$. By Proposition A.22, we may write

$$g = fq + r$$

where $r = 0$ or $0 \leq \deg r < \deg f$. By plugging in α, we see that $f(\alpha) = g(\alpha) = 0$ implies $r(\alpha) = 0$. But if nonzero $\deg r < \deg f$, violating minimality of f, so we instead have $r = 0$, implying $g = fq$ and so $f \mid g$.

To finish up, we show that f is unique. Well, if g is another monic irreducible polynomial with $g(\alpha) = 0$, then $f \mid g$ by the above argument. But f is nonzero, so $f \mid g$ requires $g = fu$ for a unit u. Being a unit means that u is a constant polynomial in K, so for example $\deg f = \deg g$, and because f and g have the same leading coefficient, we must have $u = 1$. Thus, $f = g$, as needed. ∎

Lemma A.28. Fix a field extension L/K, and let $\alpha \in L$ be algebraic over K and in particular the root of a monic irreducible polynomial $f(x) \in K[x]$. Then

$$\frac{K[x]}{(f(x))} \cong K[\alpha].$$

In particular, $K[\alpha]$ is a field of degree $\deg f$ over K.

Proof. For the first claim, note that there is a surjective ring homomorphism $\mathrm{ev}_\alpha \colon K[x] \to K[\alpha]$ by sending $g(x) \mapsto g(\alpha)$ for any $g(x) \in K[x]$. We want to show that $\ker \mathrm{ev}_\alpha = (f)$. Certainly $f \in \ker \mathrm{ev}_\alpha$. For the other inclusion, we note that any $g \in \ker \mathrm{ev}_\alpha$ has $g(\alpha) = 0$ and hence $f \mid g$ by Lemma A.27.

For the second claim, note that $K[\alpha]$ is a field because $K[x]/(f(x))$ is a field by Remark A.9. As for the degree computation, write

$$f(x) = x^d + \sum_{k=0}^{d-1} a_k x^k.$$

Then for each $n \geq d$, we can express α^n in terms of α^k with $k < n$: indeed, $f(\alpha) = 0$ implies

$$\alpha^n = -\sum_{k=0}^{d-1} a_k \alpha^{k+n-d}.$$

Thus, $1, \alpha, \alpha^2, \ldots, \alpha^{d-1}$ spans $K[\alpha]$, so $\dim_K K[\alpha] \leq d$. In fact, these α^k are linearly independent because any nontrivial relation involving them becomes a polynomial $g(x)$ with α is a root which is either zero or has degree less than $\deg f$, but Lemma A.27 enforces $g = 0$. ∎

As a last aside, we note that the sum and product of algebraic elements remains algebraic. This requires a trick known as the "determinant trick."

Proposition A.29. Fix a field extension L/K and some $\alpha \in L$. Then the following are equivalent.

(a) α is algebraic over K.

(b) The field $K[\alpha]$ is a finite extension of K.

(c) There is a subfield $K' \subseteq L$ finite over K which contains α.

Proof. We show the implications separately.

- Here, (a) implies (b) is proven in Lemma A.28.
- Note (b) implies (c) by setting $K' := K[\alpha]$.
- Checking that (c) implies (a) is harder. Suppose K' is generated by the elements $\alpha_1', \alpha_2', \dots, \alpha_n'$. Note that $\alpha\alpha_i \in A'$ for each α_i', so we may write

$$\alpha\alpha_i' = \sum_{j=1}^{n} a_{ij}\alpha_j'$$

for some elements $a_{ij} \in K'$. In other words, the matrix $T := (a_{ij})_{i,j=1}^n$ has

$$\alpha \begin{bmatrix} \alpha_1' \\ \vdots \\ \alpha_n' \end{bmatrix} = T \begin{bmatrix} \alpha_1' \\ \vdots \\ \alpha_n' \end{bmatrix}.$$

Thus, $T - \alpha I_n$ is an $n \times n$ matrix with entries in K', and it has the nonzero vector $(\alpha_1', \dots, \alpha_n')$ in its kernel, so $\det(T - \alpha I_n) = 0$. Expanding out the polynomial $\det(\alpha I_n - T) = 0$ makes α the root of a monic polynomial (of degree n) with coefficients in K', so α is indeed algebraic over K'. ∎

Corollary A.30. Fix a field extension L/K, and let K' denote the set elements of L algebraic over K. Then K' is a subfield of L. In fact, for any $\alpha \in L$ algebraic over K', we have $\alpha \in K'$.

Proof. We run our checks separately.

- We check that K' is a field. Note $0, 1 \in K'$ because these elements are the roots of the polynomials x and $x-1$, respectively. It remains to show that, for any $\alpha, \beta \in K'$, we have $\alpha+\beta, \alpha\beta \in K'$ and $\alpha/\beta \in K'$ if $\beta \neq 0$. The main point is to show that $K[\alpha, \beta]$ is a finite extension of K, which will complete the proof by Proposition A.29.

  Well, let α and β be the roots of the monic polynomials $f(x), g(x) \in K[x]$ respectively. Then by Proposition A.29 shows that $K[\beta]$ is a finite extension of K, and $f(\alpha) = 0$ shows that α is integral over $K[\beta]$, so $K[\alpha, \beta]$ is finite field extension of $K[\beta]$. We conclude $K[\alpha, \beta]$ is a finite field extension of K by Lemma A.20.

- Suppose that $\alpha \in L$ is the root of the monic polynomial $f(x) \in K'[x]$ (monic by Lemma A.27); we show that $\alpha \in K'$. Well, expand $f(x)$ as

$$f(x) = x^d + \sum_{k=0}^{d-1} a_k x^k$$

for some $a_0, \dots, a_{d-1} \in K'$. Each $a_\bullet$ is algebraic over K, so Proposition A.29 tells us that $K[a_\bullet]$ for each $a_\bullet$. As such, as in the previous check, we may build the tower

$$K \subseteq K[a_0] \subseteq K[a_0, a_1] \subseteq \cdots \subseteq K[a_0, \dots, a_{d-1}],$$

where each field is finite over the previous one by Proposition A.29. Then Lemma A.20 tells us that $K[a_0, \dots, a_{d-1}]$ is finite over K. Lastly, $f(\alpha) = 0$ tells us that α is algebraic over $K[a_0, \dots, a_{d-1}]$, so $\mathbb{Z}[a_0, \dots, a_{d-1}, \alpha]$ is finite over $K[a_0, \dots, a_{d-1}]$- and hence finite over K by Lemma A.20, meaning that α is algebraic over K by Proposition A.29. ∎

A.2.4 Enough Galois Theory to be Dangerous

We are going to derive a lot of mileage from the following result in field theory. It leads towards Galois theory; even though Galois theory is a beautiful subject, it is one that we can avoid somewhat.

Proposition A.31. Let L/K be a finite field extension, where L is a subfield of $\mathbb{C}$. Then each embedding $\sigma\colon K \to \mathbb{C}$ extends to exactly $[L:K]$ embeddings $\widetilde{\sigma}\colon L \hookrightarrow \mathbb{C}$.

Here, we are using the term "embedding" to refer to an (injective) ring homomorphism.

Proof. We induct on $[L:K]$, which is legal because $[L:K] < \infty$. If $[L:K] = 1$, then $L = K$, and there is nothing to say because we must have $\widetilde{\sigma} = \sigma$.

Otherwise, suppose $[L:K] > 1$. Then fix $\alpha \in L \setminus K$. By Lemma A.26, α is algebraic over K, so by Lemma A.27, α is the root of some monic irreducible polynomial $f(x)$. Now, $\mathbb{C}$ is algebraically closed, so we note that $f(x)$ factors as

$$f(x) = \prod_{i=1}^{n} (x - \alpha_i)$$

for some complex numbers $\alpha_1, \dots, \alpha_n \in \mathbb{C}$; note that the $\alpha_\bullet$ are distinct by Lemma A.21. Thus, given an embedding $\sigma\colon K \hookrightarrow \mathbb{C}$, there are n exactly extensions to $\sigma_i\colon K[\alpha] \hookrightarrow \mathbb{C}$ by sending $\sigma_i(\alpha) \coloneqq \alpha_i$. We have a number of checks to make this sentence make sense.

- Setting $\sigma_i(\alpha) = \alpha_i$ defines a unique embedding $K[\alpha] \hookrightarrow \mathbb{C}$. The embedding here is uniquely defined because we need to have $\sigma_i|_K = \sigma$, and then any polynomial in $K[\alpha]$ will have its output determined by where α goes. To show that σ_i is well-defined, we note that it is simply the composite

  $$K[\alpha] \cong \frac{K[x]}{(f(x))} \cong K[\alpha_i] \subseteq \mathbb{C},$$

  where the left isomorphism is by Lemma A.28.

- We have in fact defined n embeddings because the roots α_i are distinct.

- Each extension $\widetilde{\sigma}\colon K[\alpha] \to \mathbb{C}$ of σ must take this form. It suffices by our first point to check that $\widetilde{\sigma}(\alpha) = \alpha_i$ for some α_i. Well, note that

  $$f(\widetilde{\sigma}(\alpha)) = \widetilde{\sigma}(f(\alpha)) = 0$$

  because $\widetilde{\sigma}$ is a ring homomorphism. The result follows.

Now, by induction each of the σ_i extend to exactly $[L:K[\alpha]] < [L:K]$ distinct embeddings $L \hookrightarrow K$, totaling to

$$[L:K[\alpha]] \cdot [K[\alpha]:K] = [L:K]$$

embeddings $L \hookrightarrow \mathbb{C}$, where we have used Lemma A.20. Let σ_i extend to $\sigma_{i1}, \dots, \sigma_{im}$ where $m = [L:K[\alpha]]$. We have the following checks on the σ_{ij}.

- Note that σ_{ij} must be distinct: if $\sigma_{ij} = \sigma_{i'j'}$, then restricting to $K[\alpha]$ reveals that $\sigma_{ij}|_{K[\alpha]} = \sigma_i$, so $\sigma_i = \sigma_{i'}$, so $i = i'$. But then the uniqueness of extending from $K[\alpha]$ to L means that $\sigma_{ij} = \sigma_{ij'}$ implies $j = j'$.

- We show that all extensions $\widetilde{\sigma}\colon L \hookrightarrow \mathbb{C}$ of σ take the form σ_{ij}. Well, restricting $\widetilde{\sigma}$ to $K[\alpha]$ shows that $\widetilde{\sigma}|_{K[\alpha]} = \sigma_i$ for some i. Then $\widetilde{\sigma} = \sigma_{ij}$ for some j by construction of the σ_{ij}.

The above checks show that the σ_{ij} provide all extensions of $\sigma\colon K \hookrightarrow \mathbb{C}$, counted uniquely, finishing. ∎

A.2.5 Norm and Trace

Proposition A.31 allows us to make sense of the norm and trace of an algebraic element, which we now define.

Definition A.32. Let L/K be a finite extension of fields. Then for $\alpha \in L$, let $\mu_\alpha \colon L \to L$ denote the multiplication-by-α map, which is K-linear by the distributive law. Then we define the *trace* of α as $\mathrm{T}_{L/K}(\alpha) \coloneqq \operatorname{tr} \mu_\alpha$ and the *norm* of α as $\mathrm{N}_{L/K}(\alpha) \coloneqq \det \mu_\alpha$.

Example A.33. Let L/K be a finite extension of fields. Then any $\alpha \in K$ has μ_α given by the matrix $\alpha I_{[L:K]}$, so $\mathrm{T}_{L/K}(\alpha) = [L:K]\alpha$ and $\mathrm{N}_{L/K}(\alpha) = \alpha^{[L:K]}$.

Here is our key example of the norm and trace.

Proposition A.34. Let L/K be an extension of fields. Then let $\alpha \in K$ be algebraic over K which is the root of the monic irreducible polynomial $f(x) \in K[x]$. Writing $f(x) = x^d + \sum_{k=0}^{d-1} a_k x^k$, we have

$$\mathrm{T}_{K[\alpha]/K}(\alpha) = -a_{d-1} \qquad \text{and} \qquad \mathrm{N}_{K[\alpha]/K}(\alpha) = (-1)^d a_0.$$

Proof. Note $K[\alpha]$ is finite over K by Lemma A.28, where we actually showed that $1, \alpha, \alpha^2, \dots, \alpha^{d-1}$ is a basis of L as a K-vector space. Then μ_α with respect to this (ordered) basis looks like the $d \times d$ matrix

$$\begin{bmatrix} 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & \cdots & 0 & -a_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -a_{d-1} \end{bmatrix}. \tag{A.1}$$

The trace of this matrix is $-a_{d-1}$, and its determinant is $(-1)^d a_0$ by expansion by minors. ∎

Corollary A.35. Let $K/\mathbb{Q}$ be a finite extension of fields of degree n. Fix $\alpha \in K$, and let $\sigma_1, \dots, \sigma_n$ denote the embeddings $K \hookrightarrow \mathbb{C}$. Then

$$\mathrm{T}_{K/\mathbb{Q}}(\alpha) = \frac{n}{[\mathbb{Q}[\alpha] : \mathbb{Q}]} \mathrm{T}_{\mathbb{Q}[\alpha]/\mathbb{Q}}(\alpha) = \sum_{i=1}^{n} \sigma_i(n) \quad \text{and} \quad \mathrm{N}_{K/\mathbb{Q}}(\alpha) = \left(\mathrm{N}_{\mathbb{Q}[\alpha]/\mathbb{Q}}(\alpha)\right)^{n/[\mathbb{Q}[\alpha]:\mathbb{Q}]} = \prod_{i=1}^{n} \sigma_i(n).$$

Proof. We use Proposition A.34. Note $\alpha \in K$ is algebraic over $\mathbb{Q}$ by Lemma A.26, so let $\tau_1, \dots, \tau_d$ denote the embeddings $K[\alpha] \hookrightarrow \mathbb{C}$. By the proof of Proposition A.31, we see that

$$f(x) = \prod_{i=1}^{d} (x - \tau_i(\alpha)),$$

where $f(x) \in \mathbb{Q}[x]$ is the unique monic irreducible polynomial with $f(\alpha) = 0$ provided by Lemma A.27. Thus, Proposition A.34 tells us that

$$\mathrm{T}_{\mathbb{Q}[\alpha]/\mathbb{Q}}(\alpha) = \sum_{i=1}^{d} \tau_i(d) \qquad \text{and} \qquad \mathrm{N}_{\mathbb{Q}[\alpha]/\mathbb{Q}}(\alpha) = \prod_{i=1}^{d} \tau_i(\alpha).$$

To complete the proof, we must extend up from $K[\alpha]/K$ to L/K. Well, let $\ell_1, \dots, \ell_{n/d}$ denote a basis for L as a $K[\alpha]$-vector space, where we are implicitly using Lemma A.20. Then the proof of Lemma A.20 shows us that $\ell_i \alpha^j$ provides a basis for $L/\mathbb{Q}$, so writing out μ_α according to this basis looks like n/d blocks of (A.1). Because each τ_i extends to exactly n/d embeddings $K \hookrightarrow \mathbb{C}$ by Proposition A.31, the result follows. ∎

BIBLIOGRAPHY

[Old70] C. D. Olds. "The Simple Continued Fraction Expansion of e". In: *The American Mathematical Monthly* 77.9 (1970), pp. 968–974. ISSN: 00029890, 19300972. URL: http://www.jstor.org/stable/2318113 (visited on 08/26/2023).

[HW75] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*. Oxford, 1975.

[Jr02] H. W. Lenstra Jr. "Solving the Pell Equation". In: *Notices of the AMS* 49.2 (2002). URL: https://www.ams.org/notices/200202/fea-lenstra.pdf.

[Doe14] Anthony Doerr. *All the Light We Cannot See*. Scribner, 2014.

[Kle16] Felix Klein. *Elementary Mathematics from a Higher Standpoint*. Trans. by Gert Schubring. Vol. II. Springer Berlin, Heidelberg, 2016.

[Shu16] Neal Shusterman. *Scythe*. Arc of a Scythe. Simon & Schuster, 2016.

[Pro22] Ross Mathematics Program. *Students*. 2022. URL: https://rossprogram.org/students/.

[Con] Keith Conrad. *Transcendence of e*. URL: https://kconrad.math.uconn.edu/blurbs/analysis/transcendence-e.pdf.

List of Definitions