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

$$\left(x^2 - 2\right)\left(x^2 - 3\right)\left(x^2 - 6\right) = 0$$

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 $y \in \left\{\pm\sqrt{6}, \pm\sqrt{5}, \pm\sqrt{2}\right\}$, 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 $a \le b$.

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 $b \ne 0$, 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 $0 \le r < a$. 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 $0 \le r < g$. 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 $q_n := \lfloor a_n / a_{n+1} \rfloor$ if $a_{n+1} > 0$ and $(a_{n+2}, q_n) := (0, 0)$ 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 $a_0:=a$ and $a_1:=b$ and build the sequence $a_2,a_3,\ldots$ of Proposition 1.18. By induction, we see that

$$a_n\in\{a_0x+a_1y:x,y\in\mathbb{Z}\}.$$

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 $a_{n+2}=a_n-q_na_{n+1}$. 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

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

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 n*th convergent*, often denoted

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

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

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

- 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.

$$\begin{array}{cc|ccccc} & & 2 & 1 & 1 & 1 & 11 \\ \hline 0 & 1 & & & & & \\ 1 & 0 & & & & & \end{array}$$

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.

$$\begin{array}{cc|ccccc} & & 2 & 1 & 1 & 1 & 11 \\ \hline 0 & 1 & 2 & & & & \\ 1 & 0 & 1 & & & & \end{array}$$

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.

$$\begin{array}{cc|ccccc} & & 2 & 1 & 1 & 1 & 11 \\ \hline 0 & 1 & 2 & 3 & & & \\ 1 & 0 & 1 & 2 & & & \end{array}$$

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

$$\begin{array}{cc|ccccc} & & 2 & 1 & 1 & 1 & 11 \\ \hline 0 & 1 & 2 & 3 & 5 & 8 & 93 \\ 1 & 0 & 1 & 1 & 2 & 3 & 35 \end{array}$$

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.

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

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 $a_0, a_1, a_2, \ldots$ be real numbers, where $a_1, a_2, \ldots$ 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

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

for any $n \geq 0$.

Proof. The requirement that $a_1, a_2, \ldots$ 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 $(h_1, k_1) = a_1(a_0, 1) + (1, 0) = (a_1a_0 + 1, a_1)$, so

$$\frac{h_1}{k_1} = \frac{a_1a_0 + 1}{a_1} = a_0 + \frac{1}{a_1} = [a_0; a_1].$$

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

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

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

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

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 $n \ge 0$. Then for any $n \ge 0$,

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

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 $n \ge 0$, 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.

$$\begin{array}{cc|ccccc} & & 2 & 1 & 1 & 1 & 11 \\ \hline 0 & 1 & 2 & 3 & 5 & 8 & 93 \\ 1 & 0 & 1 & 1 & 2 & 3 & 35 \end{array}$$

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

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

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 $h_n/k_n := [a_0; a_1, \ldots, a_n]$ where $k_n \geq 1$ and $\gcd(h_n, k_n) = 1$. Then

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

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, we must show $k_{n+1}k_{n+2} < (k_{n+1}+k_n)(k_{n+2}-k_n)$. This rearranges to $k_n^2 < k_n(k_{n+1}+k_{n+2})$, which is true. ■

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*

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

Example 1.43. We have

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

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 1.57 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

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

Now, we claim that $k \leq \max\left\{(Cp)^{1/\varepsilon}, q\right\}$, which completes the proof. Well, suppose that $k^\varepsilon > Cp$, 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 $h_n/k_n \coloneqq [a_0; a_1, \ldots, a_n]$ 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 $\alpha_1 := (\alpha - a_0)^{-1}$, 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 $\alpha_0 := \alpha$ and

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

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

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

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

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

which completes the induction.

We now finish the proof that $\alpha = [a_0; a_1, a_2, \ldots]$. For each $n \geq 0$, set $h_n/k_n := [a_0; a_1, \ldots, a_n]$ and $h'_{n+1}/k'_{n+1} := [a_0; a_1, a_2, \ldots, a_n, \alpha_{n+1}]$ 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 Convergents Are Good Rational Approximations

As before, 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 $\gcd(h_n, k_n) = 1$. Proposition 1.40 immediately implies that

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

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

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

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

Example 1.54. Let $\varphi = \frac{1+\sqrt{5}}{2} = [1; 1, 1, \ldots]$ 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

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

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 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 $n \geq 0$, we see

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

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 1.55. Carefully tracking through Example 1.54 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 1.59.

Exercise 1.56. 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 1.57. 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

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

Simplifying, we want to show

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

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 1.58. 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 1.57, 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 1.59 (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 1.58. 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 1.57 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

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

which imply

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

By Lemma 1.39, these rearrange into

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

By Proposition 1.32, we see that $k_{3m} + k_{3m+1} \le k_{3m+2}$, so our inequalities read

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

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

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

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

1.3.4 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 1.58. Our exposition in this subsection roughly follows [HW75, Theorem 184].

> **Theorem 1.60.** 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 1.60 tells us that the best rational approximations of a real number are all continued fraction convergents.

Proof of Theorem 1.60. We use Remark 1.26 to write

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

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 1.57.) 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 $[a_0; a_1, a_2, \ldots, a_n, \ldots]$. 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 $\beta := (h' - \alpha k')/(\alpha k - h)$, 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

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

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 $m = n$. ■

1.3.5 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 (3 points). Work Exercise 1.56.

Problem 1.3.3 (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 1.3.4 (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 1.3.5 (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 1.3.6 (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 1.3.7 (2 points). Use Problem 1.3.6 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 1.54 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 1.61 (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 1.62. 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 $h := \lfloor k\alpha \rfloor$ 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 1.63. Let α be a rational number. Then $\mu(\alpha) = 1$.

Solution. Remark 1.62 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 1.64. 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 1.65. We have $\mu(\varphi) = 2$, where $\varphi = \frac{1+\sqrt{5}}{2}$.

Solution. Lemma 1.64 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 1.60, 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 1.66. Show that $\mu\left(\sqrt{2}\right) = 2$.

Remark 1.67. 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 1.68 (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 1.65 is typical, and Example 1.68 is highly remarkable.

Proposition 1.69. 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 1.64, we see that $\alpha \in S \setminus \mathbb{Q}$ must have $\mu(\alpha) > 2$. For any real number $\varepsilon > 0$, we claim that

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

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 $S_{\varepsilon,M} := S_\varepsilon \cap [-M, M]$ 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

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

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 1.65 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 1.70.** 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 1.64 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 1.71.** 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 1.70. 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 1.70 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 1.71 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 1.72. Recall from Example 1.43 that $\varphi = [1;1,1,\ldots]$. Proposition 1.71 allows us to conclude that $\mu(\alpha)=0$ immediately because $\log 1 = 0$.

Corollary 1.73. 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 1.71, 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 1.74. 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 1.65. 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 $a_{n+1} := \left\lfloor k_n^{r-2} \right\rfloor$ 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 1.71, we would like to show that

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

In fact, we claim that

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

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

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

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 1.75 (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 1.76. 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 $0 \le \deg r < \deg f$. 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 1.77 (Liouville)**.** Fix an algebraic real number $\alpha \in \mathbb{R}$ of degree $d \ge 2$. 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 $d \ge 2$ where $f(\alpha) = 0$. Namely, we may assume that f has integer coefficients by multiplying out a common denominator.

The main claim is that $|f(h/k)| \ge 1/k^d$. 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 1.76, 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, $|k^n f(h/k)| \ge 1$, 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

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

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 1.78. Fix an algebraic real number $\alpha \in \mathbb{R}$ of degree $d \ge 2$. Then $\mu(\alpha) \le d$.

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 1.77 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 1.79. By Example 1.68, the number $L := \sum_{n=0}^{\infty} 2^{-n!}$ has $\mu(L) = +\infty$. Thus, Corollary 1.78 implies that L is transcendental.

Historically, examples of the type Example 1.79 were the first numbers proven to be transcendental.

Proposition 1.77 is not at all sharp. Using bivariate polynomials instead of single polynomials, Thue was able to sharpen Proposition 1.77 into the following.

Theorem 1.80 (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 1.80 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 1.80 is still not sharp. The following result is due to Roth and is work that earned Roth the Fields medal.

Theorem 1.81 (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 1.74 with $\mu(\alpha) > 2$ were in fact transcendental! The proof of Theorem 1.81 would certainly take us too far afield, so we will not show it here.

Even though we will not prove Theorem 1.80, we will use it to the following result on Diophantine equations, as a means to reconnect with our roots.

Theorem 1.82 (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 1.80. 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 1.80.

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 1.80. ■

Example 1.83. The polynomial $f(x) := x^3 - 2$ is irreducible of degree 3. So Theorem 1.82 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 1.69 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 1.84. 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 $|q_ne - p_n| \to 0$ and $q_n \to \infty$ as $n \to \infty$. Indeed, we set $q_n := n!$ and $p_n := \lfloor q_ne \rfloor$, 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

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

so indeed $|q_ne - p_n| \to 0$ 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 $|q_np/q - p_n| \to 0$ as $n \to \infty$ by the above argument, so $|q_np - p_nq| \to 0$. However, $q_np - p_nq$ is an integer, so $q_np = p_nq$ 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 $n \mid p_nq$, 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 1.85. 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 $|q_ne - p_n| \to 0$ 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 1.86. 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 $f(x) = a_0 + a_1x + \cdots + a_dx^d$ 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 $n\ge 0$, 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

$$q_na_0+\sum_{k=1}^d a_kp_{kn}=-\sum_{k=1}^d a_k\left(q_n\alpha^k-p_{kn}\right).$$

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

$$q_na_0=-\sum_{k=1}^d a_kp_{kn}$$

for n sufficiently large. However, this cannot be: g_n divides the right-hand side for all n, so $g_n\mid q_na_0$, 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 1.87. For any polynomial f, we have

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

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 $m\ge -1$; 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 1.88. The real number e is transcendental.

Proof. Note that $e \neq 0$. We will use Lemma 1.86 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 1.87. 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

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

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 $p_{kn} := I_{f_n}(k)/(n-1)!$ and $q_n := I_{f_n}(0)/(n-1)!$ for each nonnegative integer n. We now check (a) and (b) of Lemma 1.86, 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 $g_n := n$, 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 $n(m) := 1 + md!$ and then take our rationals to be $\{p_{k,n(m)}/q_{n(m)}\}$. 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 1.89. 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 1.87, 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 $f_n(t) := t^n(t-1)^n$. Arguing as in Theorem 1.88, 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 $q_n := I_{f_n}(0)/n!$ and $p_n := I_{f_n}(1)/n!$. 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

$$J_{a,b} := \int_0^1 e^{-t} t^a (t-1)^b \, dt.$$

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

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

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 1.84), 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 $(h_n, k_n) := ((p_{n+1} + q_{n+1})/2, (p_{n+1} - q_{n+1})/2)$ 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 1.90. 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

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

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

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

Now, the fraction on the right-hand side is

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

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

Theorem 1.91. 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 1.89, 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 1.90 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 1.90 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 1.92. We have $\mu(e) = 2$.

Proof. This follows directly from plugging in Theorem 1.91 into Corollary 1.73. 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 1.66.

Problem 1.4.3 (3 points). Show that the real number

$$\sum_{k=0}^{\infty} \frac{n}{10^{n!}}$$

is transcendental.

Problem 1.4.4 (3 points). Compute

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

Problem 1.4.5 (4 points). Consider the real number

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

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

Problem 1.4.6 (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 1.4.7 (10 points). Adapt the proof of Proposition 1.89 to show that

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

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

Theme 2
Quadratic Equations

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 2.1.1 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 $f_\pm(x,y) := (\pm x + 2y, x \pm y)$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 $(x', y') := f_\pm(x,y)$ 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 $(x', y') := f_-^n(x, y)$ 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 $(x_n, y_n) := (1, 0)$ and

$$(x_{n+1}, y_{n+1}) := (3x_n + 4y_n, 2x_n + 3y_n)$$

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 $(x_0, y_0) := (2, 0)$ and

$$(x_{n+1}, y_{n+1}) := \left(\frac{x_n + 5y_n}{2}, \frac{x_{n+1} + y_{n+1}}{2} \right)$$

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, $x_n^2 + x_n y_n - y_n^2 = (-1)^n \cdot 4$ 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 $f_\pm(x, y) := \frac{1}{2}(\pm x + 5y, x \pm y)$ 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 $(x',y'):=f_-^n(x,y)$ 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 $(x_0,y_0):=(1,0)$ and
$$(x_{n+1},y_{n+1}):=(2x_n+5y_n,x_n+2y_n)$$
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, $x_n^2+x_ny_n-y_n^2=(-1)^n$ 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 two 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 $\left(2+\sqrt{3}\right)\left(2-\sqrt{3}\right) = 1$.

2.1.3 Using Continued Fractions

2.1.4 A Harder Example

2.1.5 Problems

Problem 2.1.1 (4 points)**.** 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

$$(x_{n+1}, y_{n+1}) \coloneqq (y_n, x_n + y_n)$$

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.

2.2 Quadratic Extensions

2.3 Binary Quadratic Forms

THEME 3

INTERMISSION: OTHER FIELDS

THEME 4
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]

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.

[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